Spring 2016

Pushing the boundaries: feature extraction from the lung improves pulmonary nodule classification

Samantha Kirsten Nowik Dilger
University of Iowa

PUSHING THE BOUNDARIES: FEATURE EXTRACTION FROM THE LUNG IMPROVES PULMONARY NODULE CLASSIFICATION

by

Samantha Kirsten Nowik Dilger

A thesis submitted in partial fulfillment
of the requirements for the Doctor of Philosophy
degree in Biomedical Engineering in the
Graduate College of
The University of Iowa

May 2016

Thesis Supervisor: Assistant Professor Jessica C. Sieren

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

PH.D. THESIS

_____

This is to certify that the Ph.D. thesis of

Samantha Kirsten Nowik Dilger

has been approved by the Examining Committee for
the thesis requirement for the Doctor of Philosophy degree
in Biomedical Engineering at the May 2016 graduation.

Thesis Committee: _____
Jessica C. Sieren, Thesis Supervisor

_____
Joseph M. Reinhardt

_____
John D. Newell, Jr.

_____
Eric A. Hoffman

_____
Punam K. Saha

ACKNOWLEDGEMENTS

This work would not be possible without the contributions of many. I would like to acknowledge my thesis advisor and mentor, Dr. Jessica C. Sieren. Without her guidance and support, I would not have had this incredible opportunity to continue my education, deepening my understanding of image processing, pattern recognition, and the clinical impacts. She has also been a mentor outside of research, encouraging me to explore teaching avenues, mentoring, service work, and work-life balance.

I want to express my gratitude to my committee members: Dr. Joseph M. Reinhardt, Dr. John D. Newell, Jr., Dr. Eric A. Hoffman, and Dr. Punam K. Saha. This project has benefited from your unique experiences, teachings, and areas of interest to guide the direction of this work. I would like to thank the members of the Advanced Pulmonary Physiomic Imaging Lab, including Krishna Iyer, Emily Hammond, and Abhilash Kizhakke Puliyakote for their feedback along the way.

I would specifically like to thank Deb O'Connell-Moore, Mark Escher, Allen McGruder, Dr. Michael Gailey, Monika Ahuja, Frank De Stefano, and Nicholas Koehn for their help in data collection and preparation. Thank you, Dr. Brian Smith and Sarah Bell, for your statistical support. I would like to thank Johanna Uthoff for her contribution to this project. In particular, her assistance with the clinical dataset and segmentation was invaluable. Finally, I would like to thank my husband, daughter, and family for their support and encouragement.

ABSTRACT

Lung cancer is the leading cause of cancer death in the United States. While the introduction of low-dose computed tomography (CT) screening has been shown to reduce lung cancer mortality by 20%, 97% of the suspicious lesions are found to be benign upon further investigation. Tools which can improve the specificity of CT-based screening at the time of image acquisition are desperately needed to decrease stress to the patient, avoid unnecessary radiation exposure or invasive testing, and increase cost efficiency. However, the limited amount of CT data voxels available in early identified, small (4-10mm) nodules presents a significant challenge for traditional pulmonary CT computer-aided diagnosis (CAD) tools which focus on feature extraction from only the nodule. We hypothesize a CAD tool that incorporates quantitative CT features from the surrounding lung parenchyma will improve the ability of a CAD tool to determine the malignancy of a pulmonary nodule over a CAD tool that relies solely on nodule features.

To test this hypothesis, we have developed a CAD approach that includes comprehensive intensity and texture feature extraction from the nodule and surrounding lung parenchymal tissue. We also include shape and boundary features that are minimally dependent on the preliminary nodule segmentation. Using this CAD approach, we designed two CAD tools: a research-driven lung quantification tool and clinically-focused, lung cancer screening tool robust to variations in protocol. These CAD tools were optimized for their respective datasets by exploring alternative classifiers, varying amounts of surrounding parenchyma, and the inclusion of other risk factors. In addition to use as a nodule prediction tool, preliminary clinical impact studies were performed using the optimized CAD tools by looking at alternative thresholding approaches and longitudinal nodule prediction.

For the research-driven CAD tool, a cohort of 50 (18 malignant, 32 benign) nodules from two clinical trials were identified. The optimal research CAD tool utilized the parenchyma information located within a distance of 50% of the nodule's diameter. Other valuable information came from the nodule and measures of lobar and whole lung structure. The performance of this CAD tool supports our hypothesis; a 5% improvement in accuracy was achieved with the complete feature set (95.6%), compared to the accuracy achieved with the nodule features alone (90.2%).

The clinical CAD tool was constructed with a retrospective clinical cohort of 199 nodules (109 malignant, 90 benign), divided into a 133-case training set (72 malignant, 61 benign) and a 66-case testing set (37 malignant, 29 benign). The optimized CAD tool was constructed using a least absolute shrinkage and selection operator (lasso) penalized logistic regression model and relied on nodule and parenchyma features, as well as clinical risk factors such as age and smoking history. The performance of this CAD tool was very promising, with a testing accuracy of 71% (47/66).

When three classes (definite malignant, definite benign, and indeterminate) were used, both CAD tools showed a decrease in the number of clinical follow-up procedures such as repeated imaging required (research CAD tool reduction: 30-85.7%, clinical CAD tool reduction: 12.1%). In the longitudinal trials, the CAD tools showed potential in reducing the follow-up procedures for benign nodules.

The inclusion of parenchymal features in the developed CAD tools resulted in improved performance compared to the CAD tool constructed solely with nodule features, supporting our hypothesis. We have further optimized these CAD tools by exploring the optimal amount of parenchyma for feature extraction and have demonstrated the potential of these CAD tools to positively impact clinical care.

PUBLIC ABSTRACT

Lung cancer is the leading cause of cancer death in the United States. While low-dose computed tomography (CT) screening reduces lung cancer mortality by 20%, 97% of suspicious lesions are found to be benign upon further investigation. Computer-aided diagnosis (CAD) tools can improve the accuracy of CT screening, however, current CAD tools which focus on imaging characteristics of the nodule alone are challenged by the limited data captured in small, early identified nodules. We hypothesize a CAD tool that incorporates quantitative CT features from the surrounding lung parenchyma will improve the ability of a CAD tool to determine the malignancy of a pulmonary nodule over a CAD tool that relies solely on nodule features.

Using a higher resolution research cohort and a retrospective clinical cohort, two CAD tools were developed with different intentions. The research-driven CAD tool incorporated nodule, surrounding parenchyma, and global lung measurements. Performance was improved with the inclusion of parenchyma and global features to 95.6%, compared to 90.2% when only nodule features were used. The clinically-oriented CAD tool incorporated nodule and parenchyma features and clinical risk factors and identified several features robust to CT variability, resulting in an accuracy of 71%.

This study supports our hypothesis that the inclusion of parenchymal features in the developed CAD tools resulted in improved performance compared to the CAD tool constructed solely with nodule features. Additionally, we identified the optimal amount of lung parenchyma for feature extraction and explored the potential of the CAD tools in a clinical setting.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Lung cancer is the leading cause of cancer-related death for both men and women in the United States, despite being the second-most frequent cancer diagnosis for both sexes. According to the National Cancer Institute, the age-adjusted incidence rate of lung cancer was 58.7 per 100,000 people per year, resulting in an estimated 221,200 new cases of lung cancer in 2015 [3]. While the five-year survival rate of lung cancer is low (17.4%), survival drastically improves when diagnosed in the localized stage (54.8%) [3]. This significant improvement in survival with earlier diagnosis has led to the development and implementation of screening for high-risk, asymptomatic individuals such as current and former smokers. The National Lung Screening Trial (NLST) compared two such screening techniques – projection radiograph and chest computed tomography (CT) imaging. This study demonstrated that high-risk individuals screened for lung cancer with volumetric CT imaging had a 20% relative reduction in mortality rate compared to those screened with chest radiograph [4]. This reduction in mortality is attributed to participants in the CT cohort receiving earlier stage diagnosis and treatment. However, 96.4% of the CT findings marked as suspicious for lung cancer were found to be benign upon further evaluation [5], and the improved resolution of CT has resulted in a significant increase in the number of small (4-10mm), often benign, pulmonary nodules detected. These false positives and indeterminate nodules result in unnecessary repeated CT imaging and/or invasive follow-up procedures, incurring additional emotional stress for the patient and increasing healthcare costs.

In an effort to improve the specificity of CT lung cancer screening while maintaining the improved sensitivity, a computer-aided diagnosis (CAD) tool can be designed to determine the probability of malignancy of a lung nodule based on objective measurements, or features. While many researchers have worked to develop CAD tools

for lung nodules, these CAD tools focus on properties of the nodule alone, leading to little CT data available from the voxels of small (4-10mm) nodules. Additionally, these studies neglected examination of the nodule's interaction with the surrounding lung parenchyma. We hypothesize that a CAD tool that utilizes quantitative CT (QCT) features from the surrounding lung parenchyma to determine the malignancy of a pulmonary nodule will outperform a CAD tool trained exclusively with data collected from the nodule alone. The inclusion of intensity and texture features from the lung parenchyma as well as nodule border characteristics will increase the amount of data available for classifier training without requiring additional imaging or follow up procedures for the patient.

CHAPTER II

BACKGROUND INFORMATION

2.1 Lung Cancer

While lung cancer is the second-most common cancer diagnosis of men and women in the United States, it is the leading cause of cancer death. For 2008-2012, the age-adjusted incidence of lung cancer in the US was 58.7 per 100,000 people per year; however, the age-adjusted mortality rate was 47.2 per 100,000 people per year [3]. The high rate of death relative to the incidence rate of lung cancer is strongly associated with the stage of cancer at diagnosis. While small cell lung cancer can also be staged by treatment types, both small cell and non-small cell lung cancers, such as adenocarcinomas and squamous cell carcinomas, are staged based on three criteria: the size of the tumor (T), lymph node involvement (N), and distant metastases (M) [6]. These stages describe the severity of the cancer and help guide therapy. While Stage I and II cancers are localized to the lungs alone or to the nearby lymph nodes, Stage III cancers describe more advanced disease in the chest as lymph nodes further from the initial tumor are affected. The size of the tumor is also important, with greater size leading to increasing stage. Stage IV describes the most advanced stage of lung cancer, where cancer has metastasized outside the lung [7]. The stage at which lung cancer is diagnosed affects not only treatment, but it also plays a role in patient prognosis.

The five-year relative survival rate for Stage I lung cancer is 54.8%; unfortunately, only 16% of cases are diagnosed when the cancer is in this localized stage. More than half of lung cancer cases (57%) are not diagnosed until after the primary tumor has metastasized during Stage IV disease, with a five-year survival rate of 4.2%; while 22% are diagnosed after the cancer has spread regionally in Stage II and III with a reduced five-year survival rate of 27.4%. Across all stages of lung cancer, the five-year relative survival rate is 17.4% [3]. As the majority of lung cancers are diagnosed after

metastasis, improvements in early lung cancer diagnosis are required to improve lung cancer survival.

2.2 Lung Cancer Screening

Due to the advanced stage of diagnosis when the nonspecific symptoms of lung cancer (weight loss, fatigue, cough, hemoptysis, and shortness of breath) are recognized, regular screening of asymptomatic high-risk populations, including current and former smokers, is indicated. Several types of screening have been tested including sputum testing, chest projection radiograph, and chest CT. While sputum testing and chest projection radiograph screening have not led to a reduction of mortality rates by lung cancer [8, 9], the use of low-dose CT (LDCT) as a screening tool has been shown to reduce lung cancer mortality by 20% relative to chest projection radiograph [4]. Further, in December 2013, the U.S. Preventive Services Task Force issued a statement recommending the use of LDCT for annual screening for asymptomatic adults (55-80 yrs) with a 30 pack-year smoking history and who currently smoke or have quit within the past 15 years [10]. On February 5, 2015, the Centers for Medicare & Medicaid Services (CMS) determined that, for beneficiaries meeting certain eligibility criteria, annual LDCT screening for lung cancer will be provided as an additional preventive service benefit under the Medicare program [11]. The introduction and inclusion of an insurance-covered screening program presents concern for a large increase in false-positives identified in LDCT, lesions that appear either as suspicious or too small to characterize that are found to be benign upon further follow-up. In the NLST, 96.4% of the CT findings marked as suspicious for lung cancer were found to be benign upon further evaluation [5], leading to repeated imaging tests, invasive procedures, and increased patient stress and healthcare costs.

2.3 Indeterminate Nodule Follow-up

The National Comprehensive Cancer Network® has generated an evaluation and follow-up plan for lesions identified on LDCT screening [12]. Specifically for a solid or part solid nodule identified within a LDCT screening scan, the follow-up procedure utilized is dependent on the size of the nodule (Figure 1). For nodules 6-8mm or >8mm in mean diameter, repeated imaging is often the first step in nodule evaluation, followed by biopsy or surgical resection. In repeated imaging, an increase in size is the primary criteria for intervention. Especially for these small 6-8mm nodules, a true increase in mean diameter of ≥2mm – the size increase for which further action, such as surgical resection, is recommended – can be difficult to detect, based on intra- and inter-reader variability, repeated scans, and measurement method [13, 14]. With the large increase in patients eligible for lung cancer screening and the high rate of false positives detected in LDCT, all identified nodules, but especially the smaller, sub-centimeter nodules, will continue to be a major clinical problem.

Several groups have sought to improve the specificity of LDCT by incorporating sputum testing into the screening pipeline. Burfeind Jr. et al. found the use of fluorescence in situ hybridization (FISH)-based sputum test after a nodule has been detected on LDCT to have great promise, with a positive concordance of 81.8% and a negative concordance of 91.7% to biopsy-confirmed nodules [15]. Molina et al. compared the performance of a clinical prediction model utilizing nodule size, age, and smoking status to a clinical model that also incorporated serum tumor markers. This study found an increase in performance, measured as area under the receiver operating characteristic (ROC) curve (AUC), when tumor markers were included from 0.85 to 0.93 [16]. Shen et al. incorporated microRNA biomarkers after LDCT and saw in increase in specificity to 91.8% compared to 83.6% with CT alone [17]. However, these studies require an additional sputum test to be performed in order to obtain improvement in nodule classification; while noninvasive, this test results in additional healthcare cost and

time before nodule diagnosis and subsequent treatment. Improvement in nodule classification based solely on LDCT collected information would be the optimal solution as it would not require increase in cost, additional procedures, or additional radiation exposure due to repeated imaging, and would ideally require only a minimal increase in scan processing time.

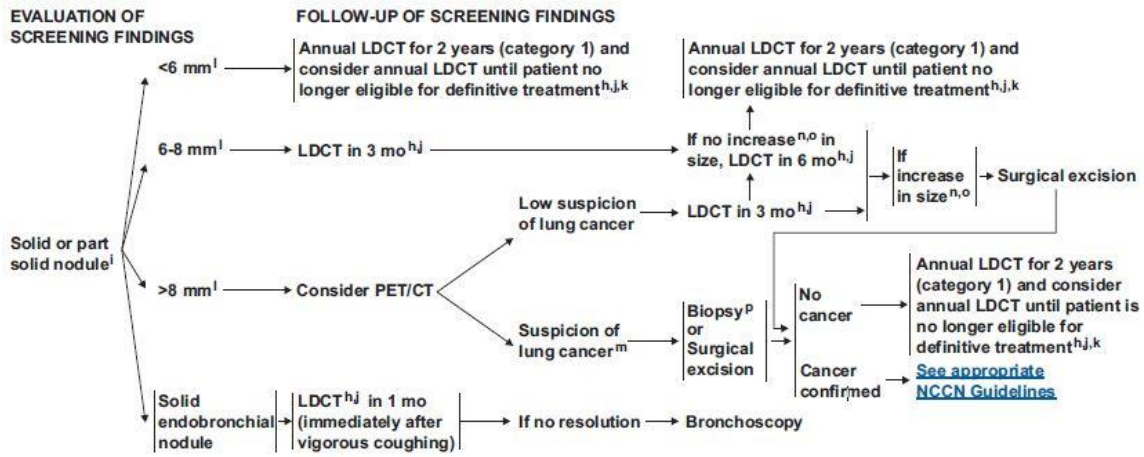

Figure 1. NCCN guidelines for the evaluation of solid or part solid nodules found during lung cancer screening. Note the frequent use of repeated imaging and/or invasive procedures to determine nodule diagnosis.

2.4 Computer-Aided Diagnosis Tools

In an effort to reduce the number of false-positives, a computer-aided diagnosis (CAD) tool can be designed to determine the probability of malignancy of a lung nodule based on objective measurements, or features. CAD tools have the potential to improve the accuracy of nodule classification (likely malignant or benign) by acting as a second reader to radiologists [18, 19]. CAD tools consist of two pathways (Figure 2). In the training pathway, features are extracted from regions of interest with known outcomes, analyzed, and selected for best classification. These selected features are then used to

train the classifier. The testing pathway takes a region of interest with an unknown outcome. The features selected in the training pathway are extracted from the unknown

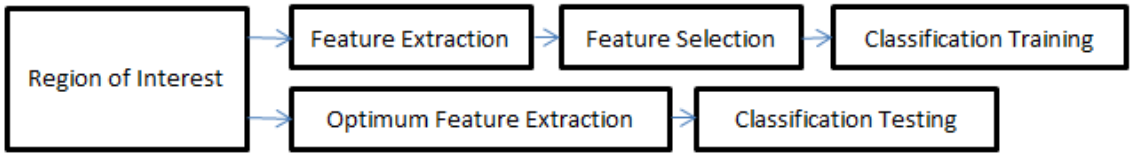

Figure 2. CAD tool pathway. The top row illustrates the training pathway which identifies the features for classification and trains the classifier. The bottom row, the testing pathway, processes new regions of interest through the trained classifier [1].

region of interest and passed to the classifier, resulting in the predicted outcome [1].

Developing a CAD tool to identify lung cancer is a topic that has been explored in many ways; however, current CAD approaches have focused on the lung nodule itself. In reviewing the current literature, size, intensity, shape, texture, and location of the nodule have all been found to be significant features in determining the diagnoses of lung nodules [19-38]. Several methods have been used to characterize the shape of the nodule including sphericity [22, 31] and compactnesss [25, 31], as well as qualitative descriptors of polygonal shape [21, 32], edge characteristics [20, 33, 34], and the presence of spiculation [19, 21, 23, 27]. The texture of the nodule has also been quantified through different techniques, such as co-occurrence matrix-based parameters, run-length statistics, and fractional Brownian motion [21, 29, 30, 36, 37]. Texture has also been determined through a radiologist scoring system [22]. While some of these features are readily computed from the regions of interest, such as attenuation properties [24, 31, 35], volume and size [25, 33, 34], and sphericity [22, 31], other features have been specified based on user input – such as the presence of spiculation [19, 21, 26, 27] and how well-defined the margins of the nodule are [23, 28]. CAD tools for lung cancer have also been

extended beyond physiological characteristics, utilizing convolutional neural networks and deep belief systems to classify pulmonary nodules [38]. The common measures of performance include accuracy, sensitivity, specificity, and AUC. Performances for these CAD tools range greatly, from accuracies of 70.0% [22] to 90.0% [30] and AUC values ranging from 0.83 [34] to 0.96 when the CAD tool was used in conjunction with a radiologist [23]. In summary, the majority of features used in classification have been derived only from the nodule, despite the fact that the nodule interacts with the surrounding parenchyma.

2.5 Project Hypothesis

The volumetric data from CT imaging facilitates the detection of very small pulmonary nodules (4-10 mm) which require follow-up, often with repeated CT imaging to track nodule growth as an indicator of malignancy. A sensitive CAD tool able to identify nodules as very low risk for malignancy would minimize repeated CT imaging and hence radiation exposure, as well as invasive procedures, in this population. However, for small pulmonary nodules, there are few CT data voxels within the solid tumor, making them difficult to process through traditional CAD tools. ***We hypothesize that improvement will be seen in the quantitative CT (QCT) CAD classification of lung nodules by including the lung parenchyma surrounding the nodule in the region of interest.*** Incorporating the surrounding lung tissue in the region of interest contributes insight into the differences between how malignant and benign nodules interact with the adjacent lung parenchyma while also increasing the amount of data (voxels) available to the CAD tool.

In developing the CAD tool, the use of CT images obtained using a diverse background of scan parameters will be explored, in an effort to identify imaging biomarkers of disease that do not rely on a strict imaging protocol. By training a CAD tool based on these parameter-invariant features, the tool will be more amenable to a

clinical setting, where it might be used on nodules identified in a screening program with a set protocol or for incidentally identified nodules during other imaging procedures.

As with reducing reliance on imaging protocol, features that reduce reliance on segmentation of the nodule are also desired. Due to the subjectivity of manual and semi-automated segmentations and differing algorithms behind automated segmentation tools, features that are robust to nodule tracings are ideal.

To explore our hypothesis we have developed three specific aims:

Aim 1: Develop and test a CAD tool that utilizes QCT features derived from the surrounding parenchyma in both a high CT resolution research cohort and a retrospective clinical cohort with greater variance in CT resolution and acquisition parameters.

Aim 2: Determine the impact of including parenchymal features by identifying the optimal amount of parenchyma surrounding a nodule to include in the CAD tool.

Aim 3: Explore the potential clinical impact of a CAD tool that includes parenchymal features in providing an early prediction of malignancy and minimizing follow-up imaging and procedures in the cohort with benign nodules.

CHAPTER III

DATASET COLLECTION

3.1 Rationale

In order to develop an accurate and effective CAD tool, a dataset with known outcomes is needed. As described above, two CAD tools are being developed: a high resolution, lung quantification tool and a low dose, lung cancer screening tool. To train these tools, two cohorts were collected. The first cohort, the high resolution, research-targeted dataset was collected from two imaging research studies specializing in lung imaging. The second cohort was collected from retrospective clinical chest CT imaging with a wide range of parameters to determine the limitations of CT variability in CAD performance.

3.2 Methods

3.2.1 Research Cohort

Two previous National Institutes of Health (NIH) funded imaging research studies were available to us for retrospective dataset collection with institutional review board approval. CT and demographic data were collected as part of these parent studies. First, the University of Iowa Hospitals and Clinics was a participating site in the National Lung Screening Trial (NLST). Hence, the Iowa cohort of NLST was the perfect starting place for developing a database of nodules with pathological diagnoses. Participants were screened annually for three years or until a malignant lung nodule was identified. Lung cancer diagnoses were tracked as a part of the primary study outcomes. By querying the Iowa cohort of the NLST, participants with either a malignant solitary pulmonary nodule (SPN) diagnosed within 3 months of an imaging timepoint or a benign SPN after all three imaging timepoints were identified.

Additionally, the Chronic Obstructive Pulmonary Disease Genetic Epidemiology (COPDGene) study is one of the largest investigations into the factors of Chronic Obstructive Pulmonary Disease (COPD) [39], with twenty-one clinical study centers across the United States. 10,000 participants were, and continue to be, imaged with CT. As there is a strong association between COPD and lung cancer [40], this study was another suitable database for potential cases. While not the primary goal of the study, an ongoing ancillary study called the COPDGene Lung Cancer Registry has been established to track participants with malignant nodules. Participants with SPNs identified from Phase I were identified. This list was cross-compared with the Lung Cancer Registry to identify participants with confirmed malignant disease. As part of Phase II of COPDGene, the participants returned for repeated imaging approximately five years after their first timepoint. For those participants with a SPN identified on their Phase I scan who were not registered with the Lung Cancer Registry, the Phase II scans were referenced to assess the SPN for 5-year stability. Nodules were assessed using the Response Evaluation Criteria in Solid Tumors (RECIST) [41]. Stability was defined as: (a) the SPN resolved and no nodule was located, (b) the SPN showed a decrease in size measured by maximum (RECIST) diameter, or (c) the SPN's maximum diameter remained constant between the two scans.

3.2.2 Clinical Cohort

With University of Iowa institutional review board approval, retrospective clinical cases were identified. Radiology and pathology reports from 2008-2014 with the phrase "pulmonary nodule" or "pulmonary nodules" were identified. These potential cases were further data-mined through the electronic health record system (Epic Systems, Verona, WI) to isolate participants meeting the following criteria: (a) radiologic presence of SPN (5-30mm) on CT and (b) true-benign or primary lung cancer confirmed by pathology or two year stability.

3.3 Results

3.3.1 Research Cohort

In total, 55 SPNs were identified from the NLST and COPDGene clinical trials (see Table A 2 for list of study identification numbers). From the Iowa cohort of the NLST, 14 cases were identified, 6 malignant and 8 benign, with diagnoses confirmed through pathology. The remaining 41 cases, 14 malignant and 27 benign, were from the COPDGene study. The malignant cases were confirmed through the COPDGene Lung Cancer Registry while the benign cases were designated benign after at least 2 years of stability on follow-up. This cohort of 55 nodules was divided into a training set consisting of 50 nodules and a testing set of 5 nodules. The training set consists of 18 malignant nodules and 32 benign nodules. The remaining five nodules, 2 malignant and 3 benign, were held separately to provide a blinded test set.

3.3.1.1 Demographics

The 20 malignant and 35 benign high CT resolution nodules acquired from 55 participants were nicely balanced in terms of demographics (see Table 1). The ages, race, proportion of females, and pack-years are consistent between the malignant and benign groups ($p>0.05$). While there were a larger percentage of participants without COPD in the benign cohort ($p<0.001$), the distribution of other stages of COPD were nearly equal. Additionally, the lobar locations of the nodules were similarly distributed between malignant and benign cases, with a preference for the upper lobes.

3.3.1.2 Nodule Size

The 50 training cases ranged in nodule diameter from 4 – 30 mm, with a median of 10.7 mm. The training cohort of malignant nodules tended to be larger as measured by radiologists (median = 13.7 mm, 4 -30 mm for malignant, median = 8.6 mm, 5 – 21 mm for benign, $p=0.005$). The 5 test cases ranged in nodule diameter from 5 – 14 mm, with a

median of 7 mm. The malignant test nodules had diameters of 12 and 14 mm, and the 3 benign test nodules ranged from 5 – 7 mm, with a median of 6.5 mm. The size summary for all 55 cases can be found in Table 1.

3.3.1.3 CT Parameters

NLST and COPDGene acquired research CT data with very different goals in mind, causing some diversity in the scan parameters as recorded in Table 2. NLST, targeting annual screening for lung cancer, developed a low-dose protocol (tube current-time product of 40-80 mAs) to minimize participant exposure to ionizing radiation. The 14 low dose CT scans from the NLST cohort were acquired from a single center site, the University of Iowa, on a Philips Mx8000, Siemens Sensation 64, or Siemens Sensation 16 CT machine. While the NLST protocol specified a slice reconstruction interval of 1.0-2.5mm, additional high resolution reconstructions ranging from 0.6-1.3mm were made at our institution and used in this study [4]. Reconstructions were performed with B (Philips) or B30f (Siemens) kernels.

The imaging protocol for COPDGene was developed for higher resolution lung imaging to best characterize emphysema, air trapping and airway wall remodeling from the COPD disease process [42]. The 41 high dose CT scans (tube current-time product of 200 mAs) from the COPDGene subcohort were collected from Brigham and Women's Hospital, Columbia University Medical Center, Morehouse School of Medicine, National Jewish Health, and the University of Iowa. Reconstructions were performed with B, B31f, B35f, or Standard kernel for Philips, Siemens, and GE respectively, with slice thicknesses between 0.6 – 0.9 mm.

3.3.2 Clinical Cohort

From 835 potential cases identified through pathology and radiology reports, 199 (109 malignant and 90 benign) were eligible given the criteria of (a) solitary lung nodule (5-30mm) and (b) true-benign or primary lung cancer confirmed by pathology or two

year stability. Further histologic breakdown of the malignant diagnosis consisted of 71 adenocarcinomas, 27 squamous cell carcinomas, and 11 other less common malignant subtypes. In the benign cohort there were 49 histoplasmosis, 14 sarcoidosis, and 27 other less common benign subtypes (Figure 3). The 199-case cohort was divided into separate training (n=133) and testing (n=66) sets with balanced diagnoses of 55% malignant and 45% benign.

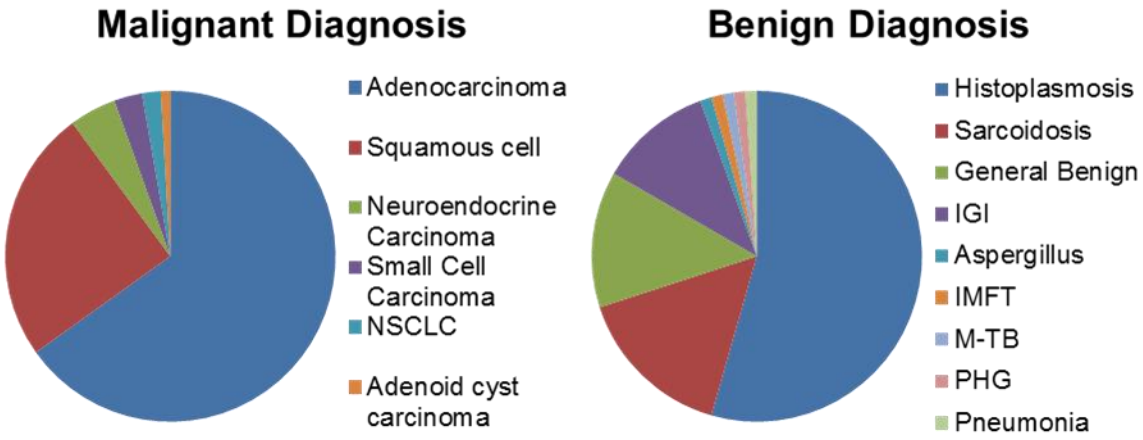

Figure 3. Distribution of malignant and benign diagnoses for the clinical cohort with lower CT resolution. Abbreviations include non-small cell lung cancer (NSCLC), indeterminate granulomatous inflammation (IGI), inflammatory myofibroblastic tumor (IMFT), mycobacterium tuberculosis (M-TB), and pulmonary hyalinizing granuloma (PHG).

3.3.2.1 Demographics

For the 199 participants, there was a statistical difference in the age ($p<0.0001$), BMI (27.3 ± 5.6, 31.8 ± 9.0, $p=0.00034$), sex ($p=0.0034$), and pack-years ($p<0.0001$) between malignant and benign cases. COPD Gold stage was also statistically significant between the malignant and benign cases ($p<0.0001$), with a larger percentage of participants without COPD in the benign group. Nodule lobar location was found to be

significant (p=0.04), with a larger percentage of malignant nodules in the upper lobes compared to the benign nodules. Race was not statistically significant between the two classes. These characteristics are summarized in Table 1.

3.3.2.2 Nodule Size

There was also a statistically significant difference between the size of the nodules in each diagnosis group (Table 1, $p<0.0001$). The malignant nodules had an average maximum diameter of 18.5±2.4mm, whereas the benign nodules had an average maximum diameter of 13.7±6.0mm. The primary reason for this size disparity rises from the method of CT scan collection. To ensure the most accurate diagnostic information, the CT scan closest to the time of histologic confirmation with the nodule size remaining under 30mm was selected. Due to malignant nodules' more aggressive growth rate, it is likely the scan selection procedure led to a bias in the size of the nodules.

3.3.2.3 CT Parameters

A variety of scanner protocols were utilized in the 199 acquired cases (Table 2). Four CT manufacturers (Siemens (162), GE (15), Philips (2), and Toshiba (20)) and 18 CT models were used. The most common models were: Siemens SOMATOM Definition (71), Siemens Sensation 16 (46), Sensation Biograph 40 (15), and Toshiba Aquilion (20). The reconstruction kernels were divided categorically into standard (163), soft (2), sharp (7), standard without beam hardening correction (11), iterative (6), and Toshiba standard (10) (detailed representation of kernel types in Table A 1). Using these machines, several protocols were used, including 115 Chest CT with and without contrast, 20 CT: Chest, Abdomen, Pelvis, 7 CT Angiography, 45 External CT scans, and 12 PET/CT scans. The scan parameters had an average tube voltage of 118 kVp and average tube current of 414 mA. Slice thickness ranged from 1.0-6.0mm (average of 3.3mm), and 184 scans used contrast. There was no statistical difference between diagnosis groups for tube voltage, current, or slice thickness.

Table 1. Comparison of demographics between research CT (high resolution) and clinical CT (lower resolution) cohorts. Similar age distributions and pack years can be seen across the two cohorts, with both benign groups having fewer pack-years than the malignant groups. The benign groups also have a greater proportion of individuals without COPD. Also noted is the increased nodule size in the malignant groups.

	Research CT (55)		**Clinical CT (199)**	
	20 Malignant	35 Benign	109 Malignant	90 Benign
Age	65±7 yrs	65±8 yrs	66±11 yrs	53±13 yrs
Pack-years	53±34 pk-yrs	50±26 pk-yrs	38±32 pk-yrs	11±17 pk-yrs
Men:Women	9:11	15:20	60:49	28:62
% without COPD	20%	48.6%	17.4%	38.9%
Nodule Size Range, Median	4-30mm, 13.7mm	5-21mm, 8.6mm	6-30mm, 20.0mm	6-30mm, 13.0mm
% nodules in upper lobes	55%	54.3%	60.5%	43.3%

Table 2. Comparison of scanner parameters between research CT (high resolution) and clinical CT (lower resolution) cohorts. While similar kilovoltage and current averages are seen, there is greater diversity seen in these metrics in the clinical cohort. Additionally, the clinical CT scans are much thicker than the research CT scans.

	Research CT (55)	**Clinical CT (199)**
Kilovoltage Range, Mean	120-120 kVp, 120 kVp	80-140 kVp, 118 kVp
Current Range, Mean	50-440 mA, 340 mA	36-795 mA, 413 mA
Slice Thickness Range, Mean	0.6-1.3mm, 0.76mm	1.0-6.0mm, 3.3mm
Most common CT machines and kernels	Siemens: B31f (38) Siemens: B30f (11) Philips: B (3) GE: Standard (2) Siemens: B35f (1)	Siemens: B31f (119) Siemens: B30f (14) Siemen: B31s (10) GE: Standard (10) GE: Lung (10)
% with iodine contrast	0%	92%

3.4 Discussion and Conclusions

With these two cohorts, the two goals of this work – a dedicated research, lung quantification tool and a clinically-useful nodule prediction tool – can be designed. For the research cohort, a collection of low and high dose CT scans with thin slice reconstructions was acquired to provide a more uniform dataset while still including opportunities for variation in scan protocol for robust feature identification.

Although an ideal clinical cohort would contain non-contrast, low dose CT scans only, as this is the protocol that was approved for lung cancer screening [10], the Iowa lung cancer screening program is still in development. As a surrogate, we sought to collect scans with wide variability in scan parameters, such as voltage, exposure, reconstruction kernel, and scanner models and manufacturers, in order to identify features that are more robust to protocol variability.

CHAPTER IV

CAD DEVELOPMENT

Computer-aided diagnosis (CAD) tools use a training set with known outcomes or diagnoses to select a set of pertinent features, as described in *2.4 Computer-Aided Diagnosis Tools*. From the features in the training set, a classifier learns to generate a prediction for a new case of unknown outcome or diagnosis. In this chapter, the development of the research-driven, lung quantification tool is described in *4.1 Research Cohort*. The modifications to the research-driven CAD tool for application to the clinical cohort are described in *4.2 Clinical Cohort*.

4.1 Research Cohort

Described in *3.3.1 Research Cohort*, the research cohort consisted of 50 training and 5 test subjects. The 50 training cases were used to develop the CAD modules shown in Figure 4, including: Segmentation, Feature Extraction, Feature Selection, and Classification.

4.1.1 Methods

For the research cohort with high CT resolution, each of the 50 training and 5 test subjects had a region of interest (ROI) defined (Figure 4, Segmentation). The ROI was selected within the CT data to include the solitary pulmonary nodule and immediate surrounding parenchyma. The amount of parenchyma included in the ROI was roughly proportional to the size of the nodule. Each ROI was manually segmented by a single user (J.U.) to label the individual voxels of the ROI as nodule, surrounding parenchyma, or invalid tissue. Voxels marked as invalid tissue included chest wall and blood vessels; these voxels were not used in feature extraction.

4.1.1.1 Feature Extraction

Voxels labeled as parenchyma and nodule were used in the extraction of four classes of features (Figure 4, Feature Extraction): intensity, shape, border, and texture. Histogram features such as mean, minimum, and maximum intensity (in Hounsfield Units, HU), as well as measures of heterogeneity such as entropy and kurtosis were extracted from the nodule and parenchyma voxels separately. Shape characteristics such as sphericity and features comparing the surface of the nodule to a sphere of equivalent volume were extracted using the nodule mask. The size of the nodule was also calculated using the RECIST maximum diameter [41] and the effective radius, the radius of a sphere of equivalent volume of the nodule.

These four classes of features are described in greater detail in [1]; however, some key features unique to this CAD approach are summarized here, including texture and border features. For the texture features, 272 three-dimensional (3D) texture features were extracted from the nodule and parenchyma using Laws' Texture Energy Measures (TEM) [43]. Laws' TEMs interrogate the grey-scale images by looking for patterns in different gradients. In two-dimensional (2D) applications, five 5-element vectors describing levels, edges, spots, waves, and ripples are convolved to form 25 2D 5x5 kernels. These kernels locate a different texture type in each dimension. By convolving the original 5-element vectors with the 2D kernels, 125 3D kernels were generated. These 3D kernels were convolved with each ROI to create texture volumes which were normalized and combined to form 34 rotationally-invariant 3D TEMs [2]. From these TEMs, the mean, variance, kurtosis, and skewness of the nodule and parenchyma were extracted.

The border was analyzed by extending the rubber band straightening transform (RBST), introduced by Sahiner [44], so the nodule is straightened from its centroid. This shift in the straightening origin provides less reliance on the segmentation of the nodule, as the nodule's centroid is more consistent across varying segmentations. This is

illustrated with 12 nodules segmented by three users in [2], where the mean centroid difference was 0.45±0.41mm. The axial slice of the ROI containing the centroid was selected, and the distance between the centroid and the closest bounding box edge was computed and determined to be the ray length. Rays were cast from the centroid toward the edge of the bounding box radially, resulting in 360 rays of uniform length, each representing a single row in the straightened image. From the straightened image, the nodule's border definition was assessed based on the intensity rate of change along the nodule-parenchyma interface [2]. A list of equations for less common features can be found in Table A 3, and a full list of features extracted by the research CAD tool can be found in Table A 4.

4.1.1.2 Feature Selection and Classification

In total, the feature extraction module computes 304 nodule and nodule-associated parenchymal features from the ROI. These features were then reduced through statistical analysis and feature selection (Figure 4, *Classification*). A statistical analysis module was created to determine which features were statistically different between the malignant and benign cases. Each feature was tested for normality using the Jarque-Bera normality test. Additionally, the variances for each feature were tested for equality. If the feature followed a normal distribution and the variances for the malignant and benign groups were equal, a two-sample t-test for equal variances was performed. If the feature followed a normal distribution but did not have equal variances between the two groups, a two-sample unequal variances t-test was used. Finally, if the normality test showed the feature distribution was not normal, the nonparametric Wilcoxon Rank Sum test was used. A feature with a p-value less than 0.05 was determined to be statistically significant [1, 2]. This statistical testing was used to reduce the number of features used in feature selection in order to reduce computation time while ensuring meaningful features were included in analysis.

To determine which of the significant features should be used for classification, stepwise forward selection was used. A framework was developed for optimizing feature selection when a small number of ROIs were available in an effort to reduce both performance and selection biases. The performance of the selected features at each step was estimated using leave-one-case-out cross-validation on 49 cases, with the 50th case completely left out of training and testing (Figure 5). This was repeated so every case was completely removed from the process. Majority voting was then used to select the best feature set at each stage of stepwise forward selection [2]. Features were added until the best combination of features is found, with the total number of features equivalent to 1 feature for every 10 regions of interest to prevent over-fitting [45]; with 50 ROIs, features were added until 5 features were selected or until performance was not improved with the addition of more features. An artificial neural network (ANN), the structure of which is illustrated in (Figure 4, Classification) was used for classification, using a leave-one-out cross-validation method to estimate performance. Due to the random initialization of weights during training of the ANN, this was repeated ten times to better approximate the classifier's performance. Classification using the features selected was also performed using a linear discriminant analysis (LDA) classifier; due to the deterministic nature of LDA, the leave-one-out cross-validation method was performed once [2].

Four feature selection and classification trials were designed to investigate the impact of including different types of features and classifiers (Figure 6). These trials include: nodule and parenchyma features (N+P) and nodule only features (N) with an ANN and a LDA classifier.

4.1.1.3 Independent Test Cohort

In addition to testing with a leave-one-case-out method, the developed classification approaches were tested using the independent test dataset consisting 5 COPDGene cases (2 malignant and 3 benign). As these cases were collected after training

and validation of the first 50 cases had been performed, this dataset represented a truly blinded testing set with which to ensure overfitting had not occurred.

4.1.1.4 Classification Evaluation

The accuracy, sensitivity, specificity, and the standard deviations of each measure were computed by averaging the leave-one-case-out cross-validation results from 10 trials with each ANN classifier. For both ANN and LDA classifiers, performance was evaluated through ROC analysis in which the area under the ROC curve (AUC) was obtained. Additional performance measures of positive predictive value (PPV) and negative predictive value (NPV) were also computed. Classifier ROC curves were compared using the Delong method to assess for differences in performances [46]. Predicted classification probabilities were averaged over the 10 trials for comparisons of ANN performance. All statistical comparisons were two-sided and assessed for significance at the 5% level.

4.1.2 Results

4.1.2.1 Statistical Analysis of Features

Statistical analysis of the 304 features (159 nodule, 145 parenchymal) found 53 statistically significant predictors of malignancy ($p<0.05$). In summary, the significant predictors incorporated features from both the nodule and the surrounding parenchyma:

- Eleven intensity features - nodule mean, median, and minimum intensities, nodule variance, nodule kurtosis, nodule skewness, parenchyma mean and median intensities, parenchyma variance, parenchyma entropy, and parenchyma kurtosis.
- Two shape features – sphericity (the ratio of surface area to the volume of the nodule) and the range of the variation between the nodule's boundary and the sphere of equivalent volume.

- Two size measures – the effective radius (the radius of a sphere of equivalent volume) and the RECIST maximum diameter [32].
- Four border features – the mean and standard deviation of the border's width in physical space, the mean of the border's width in pixel space, and the mean intensity change across the border.
- Thirty-four 3D nodule texture kurtosis features.

Additionally, 26 parenchymal texture variance features were near significant ($p<0.1$).

Comparatively, when only the nodule features were considered, 49 features returned as significant:

- Six intensity features – nodule mean, median, and minimum intensities, nodule variance, nodule kurtosis, and nodule skewness.
- Two shape features – sphericity and the range of variation between the nodule's boundary and the sphere of equivalent volume.
- Two size measures – the effective radius and the RECIST maximum diameter.
- Five border features – the mean and standard deviation of the border's width in physical and pixel space and the mean intensity change across the border.
- Thirty-four nodule texture kurtosis features.

4.1.2.2 Classification

We ran four trials to test our hypothesis and found that utilizing QCT features from the surrounding parenchyma improve performance of a CAD tool in determining the malignancy of a pulmonary nodule. The summary of features selected by our modified stepwise feature selection approach and classifier performances of these four trials (ANN and LDA with N and N+P features) can be seen in Table 3 and Table 4.

From the subset of 53 statistically significant nodule and parenchyma features, the features chosen by stepwise forward selection for the N+P ANN classifier are shown in Table 3, along with the performance of the classifier using leave-one-case-out cross-validation. Likewise, the features selected for classification from the 49-feature subset for the nodule only, N, are also shown in Table 3. For the N+P ANN classifier, only 4 features were selected when up to 5 could be used to maintain the 1 feature per 10 cases rule of thumb. In this instance, performance was not improved with the addition of a fifth feature. In both the N and N+P trials, the nodule's texture kurtosis played a major role in classification. Interestingly, in the N+P classifier, the parenchymal variance of HU was selected first, outperforming all other features used independently.

Box plots for these four features selected by the N+P ANN are shown in Figure 8. In the nodule's texture kurtosis features, larger variances are observed in the malignant nodules. Similarly, the malignant nodules had greater variance in parenchyma intensity values; this was expected due to spiculation and angiogenesis in the surrounding parenchyma. Finally, while the malignant nodules had more dense minimum intensities, the variance in HU was larger in benign nodules.

Inclusion of parenchyma features improved classifier performance in both the ANN and LDA over classifiers utilizing nodule features alone; comparisons of the ROC curves for these classification methods are shown in Figure 7. Looking at the ANN classifiers, the classifier incorporating parenchyma features in addition to nodule features resulted in an AUC of 0.913 whereas the classifier constructed solely from nodule features had an AUC of 0.882. This difference is not statistically significant ($p=0.55$), although the modest sample size limits power to detect small difference. Similarly, in the LDA comparison, the N+P classifier outperformed the N classifier (0.821 vs 0.574, $p<0.01$). For both N and N+P classifiers, the AUCs were significantly higher for the ANN classifiers compared to the LDA classifiers (N: $p<0.01$, N+P: $p<0.01$). Testing on the independent cohort showed the classifiers did not suffer from overfitting. The N+P

ANN had an accuracy of 80% (4/5), sensitivity of 100% (2/2), and specificity of 67% (2/3). The N classifier had even better blinded performance, with an accuracy of 96% (4.8/5), 100% sensitivity, and a specificity of 93% (2.8/3). The LDA blinded tests performed worse with an accuracy (sensitivity, specificity) of 60% (50%, 67%) for the N+P LDA and 60% (0%, 100%) for the N LDA. Additional performance metrics can be seen in Table 4.

4.1.3 Conclusion

Existing CAD tools have focused primarily on characteristics derived from the nodule and may be challenged by early detected lesions due to the limited number of CT voxels within the lung nodule. To overcome this obstacle, nodule-associated parenchymal features were incorporated. By utilizing these nodule and parenchyma features and by incorporating data collected using a dataset with a diverse range of scanning protocols, we have developed a robust classification tool. Our comparison of classifiers found that the ANN trained with both nodule and parenchymal features trended toward increased performance in accuracy, sensitivity, PPV, NPV, and AUC when compared to performance using nodule feature alone despite reliance on fewer features. This N+P classifier relied on nodule minimum intensity, parenchyma variance in intensity, and two 3D nodule kurtosis texture features. The inclusion of parenchyma features and improved performance of the CAD tool supports our hypothesis that valuable information is present within the parenchyma and can be captured from parenchymal features.

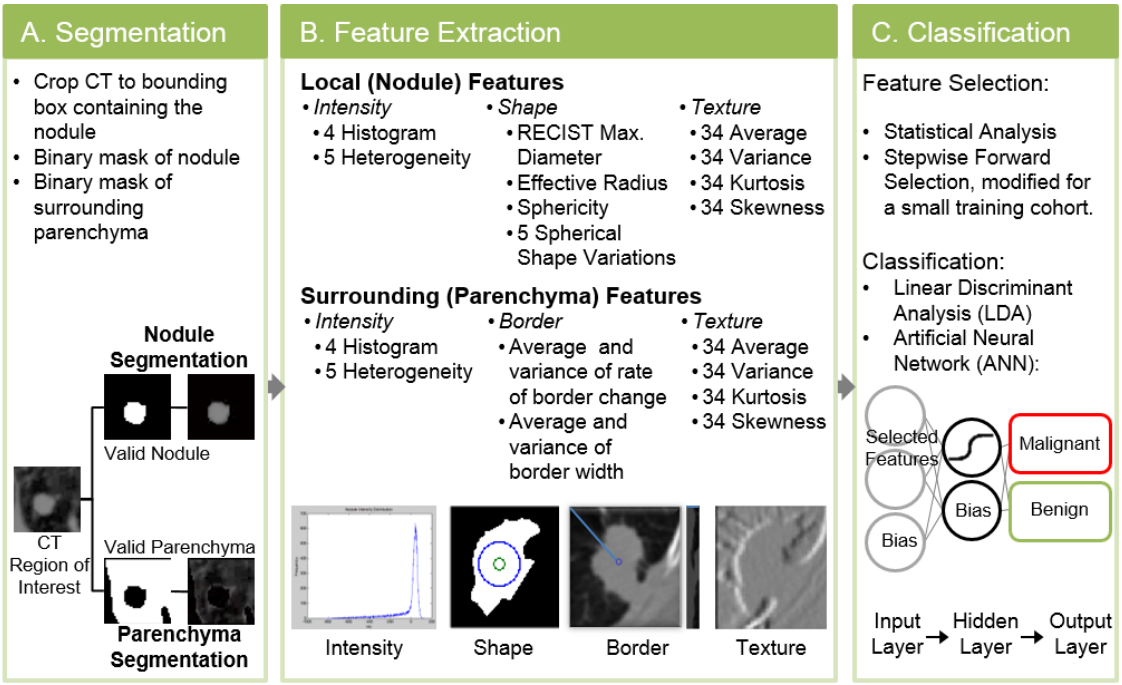

Figure 4. Flowchart of the computer-aided diagnosis tool. The region of interest containing the nodule and surrounding parenchyma is labeled as nodule or parenchyma (A. Segmentation). In B. Feature Extraction, over 300 features are extracted from these two areas. Finally, the feature set is reduced and the selected features are used to train the neural network classifier (C. Classification), labeling a nodule as likely malignant or likely benign.

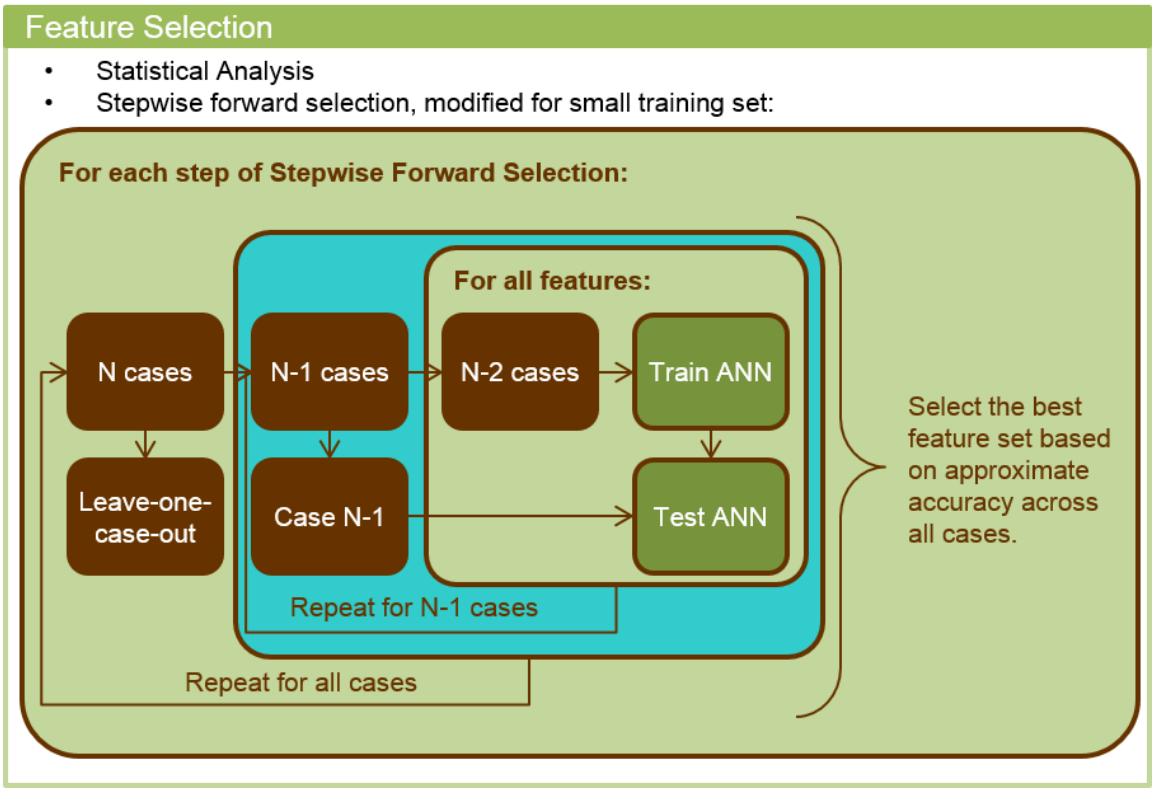

Figure 5. Feature Selection protocol. After a first-wave statistical analysis to identify statistically significant features, the modified Stepwise forward selection with majority voting was used to select up to floor(N/10) features, where N is the total number of cases in the cohort. For each iteration of stepwise forward selection, one case is left completely out of the internal loop, where accuracy approximations are computed for each subset of features being tested. After acquiring accuracy approximations for all possible subsets for each set of 49 (N-1) subcohorts, the feature set with the highest approximate accuracy across all subcohorts is selected [2].

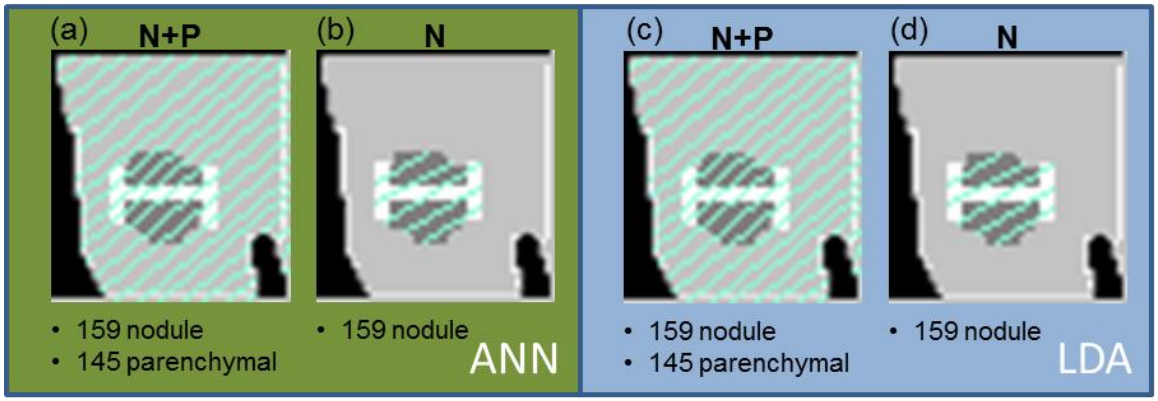

Figure 6. Four trials exploring the impact of various regions and classifiers were designed. (a) N+P builds an ANN classifier from nodule (N) and parenchyma features (P) , whereas (b) builds an ANN classifier from nodule features (N) alone. (c) and (d) explores the impact of using a LDA classifier on the N+P and N feature sets.

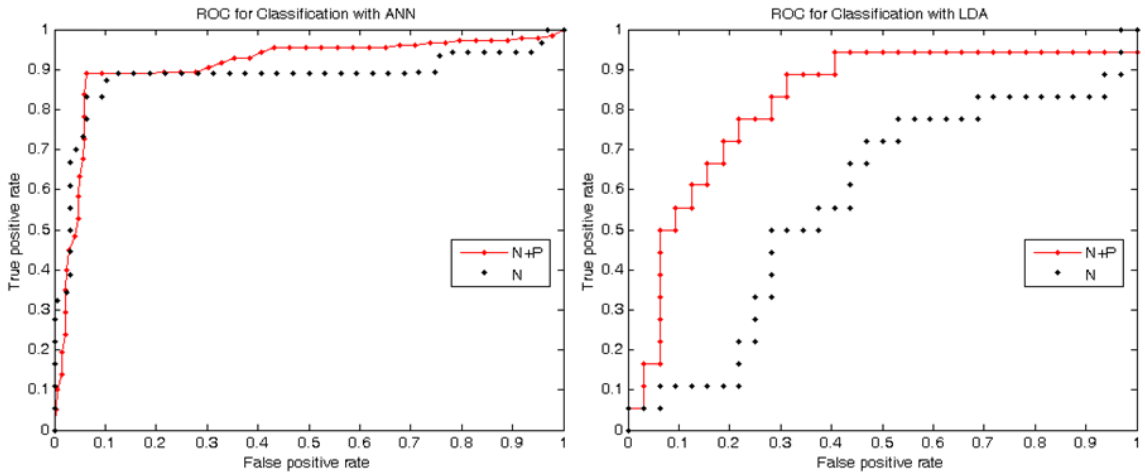

Figure 7. ROC curves for the ANN (a) and LDA (b) classifiers. The ROC curves generated from the ANN classifiers show better performance compared to the LDA classifiers. This is supported by the AUC values seen in Table 3 and Table 4.

Table 3. Summary of performance measures for the artificial neural network (ANN) classifications. Due to the random initialization of the ANNs during training, performance is reported in percent as the mean ± the standard deviation over the 10 trials. The positive predictive value (PPV), negative predictive value (NPV) and the area under the ROC curve (AUC) are also reported.

Selected Features	ANN Leave-one-case-out Classifier Performance						
	# correct 50 (18, 32)	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Nodule & Parenchyma Features: N+P Parenchyma Variance Nodule Minimum HU Nodule Kurtosis TEM1 Nodule Kurtosis TEM15	45.8 (16,29.8)	91.6±0.8%	88.9±0.0%	93.1±1.3%	.879	.937	0.913±0.019
Nodule Features: N Nodule Variance Nodule Kurtosis TEM17 Mean of Border Nodule Kurtosis TEM2 Nodule Kurtosis TEM16	44.9 (15,29.9)	89.8±0.6%	83.3±0.0%	93.4±1.0%	.877	.908	0.882±0.005

Table 4. Summary of performance measures for the linear discriminant analysis (LDA) classifications. The accuracy, sensitivity, specificity, PPV, NPV, and the AUC are reported.

Selected Features	LDA Leave-one-case-out Classifier Performance						
	# correct 50 (18, 32)	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Nodule & Parenchyma Features: N+P Parenchyma Variance Nodule Minimum HU Nodule Kurtosis TEM1 Nodule Kurtosis TEM15	38 (8,30)	82.1%	76.0%	93.8%	.800	.750	0.821
Nodule Features: N Nodule Variance Nodule Kurtosis TEM17 Mean of Border Nodule Kurtosis TEM2 Nodule Kurtosis TEM16	29 (2,27)	58.0%	11.1%	84.4%	.286	.628	0.575

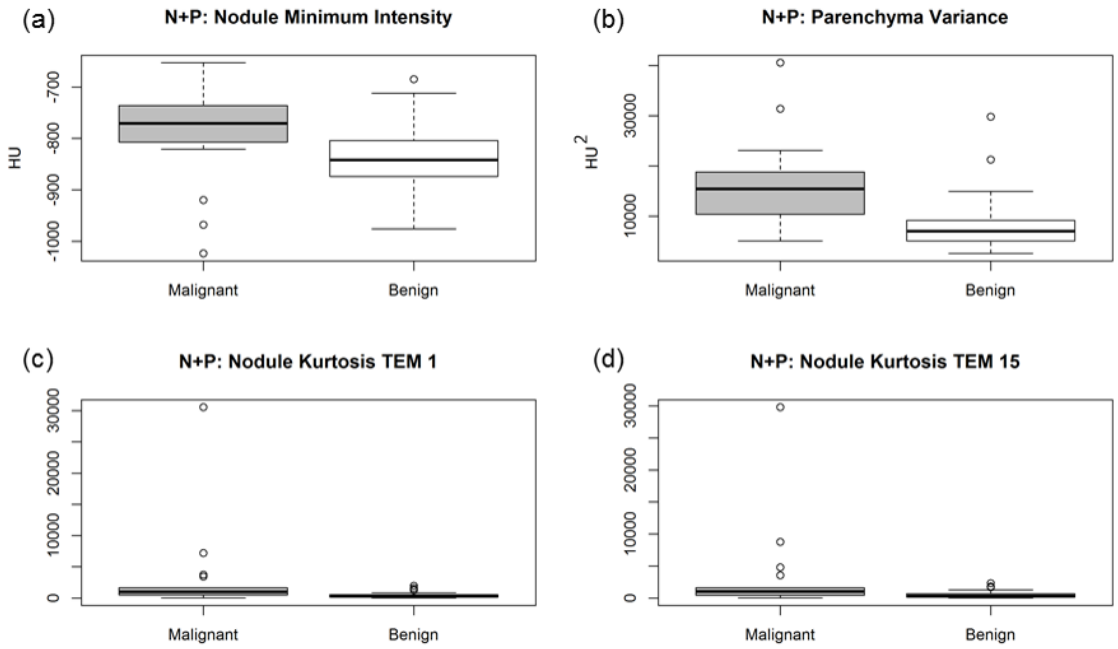

Figure 8. Box plots of features selected by the research CAD tool constructed using nodule and parenchyma features (N+P). Four features were selected: (a) nodule minimum intensity, (b) parenchyma variance, and (c-d) 2 3D nodule kurtosis texture features.

4.2 Clinical Cohort

4.2.1 Rationale

With the increase in classifier performance in the research cohort when nodule and parenchyma features were utilized, we sought to modify the research-driven lung quantification CAD tool in order to develop our clinically-focused, lung cancer screening tool. Our clinical cohort was much larger (n=199 vs. n=50) and more varied in terms of CT scanner protocol (Table 2). Additionally, the clinical cohort had much thicker CT slice thicknesses, causing reduced resolution. To develop a robust clinical CAD tool, several modifications were implemented in the feature extraction, feature selection, and classification modules of the CAD tool (Figure 9).

4.2.2 Methods

As discussed in *3.3.2 Clinical Cohort*, the 199 retrospectively collected cases were randomly divided into a 133-case training set and a 66-case testing set. The training set was used for feature selection and training of the classifiers, and the testing set was used as a blind test sample to gauge and compare performances of the trained classifiers.

4.2.2.1 Feature Extraction

The research cohort was collected from two tightly-controlled clinical trials. As a result, the ages, sexes, and smoking histories of the malignant and benign cases were similar. In the clinical cohort, there is much greater inter- and intra- class variability in risk factors such as smoking history and age (Table 1). As a result, trials incorporating these clinical risk factors were performed (N+P+R), as well as trials exploring the impact of the nodule and parenchyma QCT features alone (N+P).

Due to the variability of slice thickness within the cohort (range: 1.0-6.0mm, median: 3.0mm), we hypothesized the two-dimensional implementation of Laws' TEM would perform better than the three-dimensional Laws' TEMs. While the 3D Laws' TEMs were significant in the research cohort, the slice thicknesses were more uniform, providing similar volumes over which to compute the textures. In the clinical cohort, the 3D kernels are applied to a variety of voxel sizes due to the variance in the z-plane (1.0-6.0mm), and thus increasing the amount of both intra- and inter- class texture variability. To better isolate inter-class variability, the 2D Laws' TEMs were applied to the centroid slice of the nodule's ROI. The range in voxel size in the x- and y- planes was 0.55-0.98mm, illustrating a more standardized area over which to compute texture. The centroid was computed from the segmentation during border analysis with the RBST; the slice containing the centroid was also identified and Laws' 2D TEMs were extracted from it in a similar manner as described in *4.1.1.1 Feature Extraction*. From both the nodule and parenchyma regions, the mean, variance, kurtosis, and skewness for each of

the 14 2D TEMs were computed, resulting in the computation of 112 texture features. A complete list of features can be found in Table A 5.

4.2.2.2 Feature Selection

In the research cohort, a leave-one-case-out majority voting framework was developed to select the features for use in classification in a way that minimized the amount of bias introduced while allowing all cases to be used for classification. While valuable for the smaller research cohort of 50 cases, this process was too time-intensive to be performed in the larger, more diverse clinical cohort. The ability to divide the clinical cohort into training and testing sets also made the majority voting framework unnecessary.

Collaborations with colleagues (Sarah Bell and Dr. Brian Smith) in the Department of Biostatistics led to a least absolute shrinkage and selection operator (lasso) penalized logistic regression model (R, www.r-project.org) method to select features [47]. In lasso-penalized logistic regression, the number of variables is limited by the introduction of a λ penalty term. The larger the value of λ, the more variables allowed in the model; the smaller the λ value, the fewer variables selected. Essentially, λ controls how many of the variables in a standard logistic regression formula have coefficients of zero. To determine the value of λ that provided the highest average AUC, 1,000 iterations of 10-fold cross-validation were employed using the training set. As the selection of this λ determined which features received zero coefficients, the features with non-zero coefficients are those selected for classification.

4.2.2.3 Classification

The features selected through lasso-penalized logistic regression (LR) were also implemented with ANN and LDA classifiers. These three classifiers (LR, ANN, and LDA) were trained using the training set, and performance was evaluated using the blinded testing set.

Two trials were performed: N+P, in which the CAD relied solely on the nodule and parenchyma QCT features, and N+P+R, which included risk factors such as age, sex, smoking history, and nodule lobe location in addition to QCT features. These two trials processed through the three classifiers resulted in six models (Figure 10). Performance was assessed through accuracy, sensitivity, specificity, PPV, NPV, and AUC of the testing set. The AUCs of the training set were compared to the testing set to ensure overfitting had not occurred to maintain generalizability.

4.2.3 Results

Six models were developed while testing the CAD performance on the clinical cohort: N+P for LR, ANN, and LDA, and N+P+R for LR, ANN, and LDA (Figure 10). Selected features and summaries of classifier performances can be seen in Table 5 and Table 6. In both N+P and N+P+R trials, the LR classifier performed better than the other classifiers. Additionally, the selection of parenchyma features in both N+P and N+P+R trials illustrate the inclusion of parenchyma features adds value to the CAD tool.

4.2.3.1 Feature Selection

Prior to feature selection, the scan parameters including contrast use, slice thickness, voltage, current, and exposure were examined for statistical differences between malignant and benign cases. No significant association was present between these parameters and the nodules' diagnoses.

Using lasso-penalized logistic regression, parenchyma features contributed greatly. When risk factors were incorporated, the lobe location, age, sex, and pack years played a significant role in classification. In both N+P+R and N+P trials, nodule entropy, parenchyma mean intensity, and 2D nodule skewness textures were selected. The box plots for these features between malignant and benign nodules are shown in Figure 11. Of importance, the parenchyma mean intensity is less dense for malignant nodules compared to benign nodules. This could suggest the presence of lung diseases such as emphysema,

which is associated with lung cancer. When QCT features alone were used, the parenchyma played a more prominent role, with both the parenchyma mean and minimum intensity being selected.

4.2.3.2 Classification

The classifier performances can be seen in Table 6. Models constructed with risk factors (N+P+R) outperformed those constructed from QCT features alone (N+P). Across classifiers, the performances were very similar; however, the models constructed with the LR classifiers slightly outperformed the others (N+P testing AUC: 0.68 (LR) vs. 0.66 (ANN), 0.68 (LDA); N+P+R testing AUC: 0.81 (LR) vs. 0.76 (ANN), 0.76 (LDA)).

4.2.4 Conclusion

The best overall model was the N+P+R LR model, which incorporated nodule and parenchyma intensity measures, shape characteristics, and nodule textures. The accuracy of this model on the testing cohort was 71% (47/66) with an AUC of 0.81. An additional benefit of this model was the lack of size as a feature. As a size bias is present in our clinical dataset, the lack of size as a feature is an indicator that several stronger features better separate the classes. In general, the N+P+R trials performed better compared to the N+P trials. However, evidence of overfitting is present for the ANN N+P+R model. The training AUC value was nearly perfect (0.97) while the testing AUC was substantially lower (0.76).

The best model constructed from QCT features only was the N+P LR model with an AUC of 0.68. This CAD tool incorporated the size feature of effective radius. When size was eliminated as a feature, the training AUC dropped slightly to 0.76 (from 0.79 with size), and the testing AUC also dropped from 0.69 with size to 0.58 without size features (Table A 8). Additional N+P and N+P+R classifiers constructed with different feature subgroups (2D vs. 3D textures) were performed (Table A 7 and Table A 8).

Similar performance results were achieved for N+P and N+P+R using 2D textures vs. 2D and 3D textures or 3D textures alone.

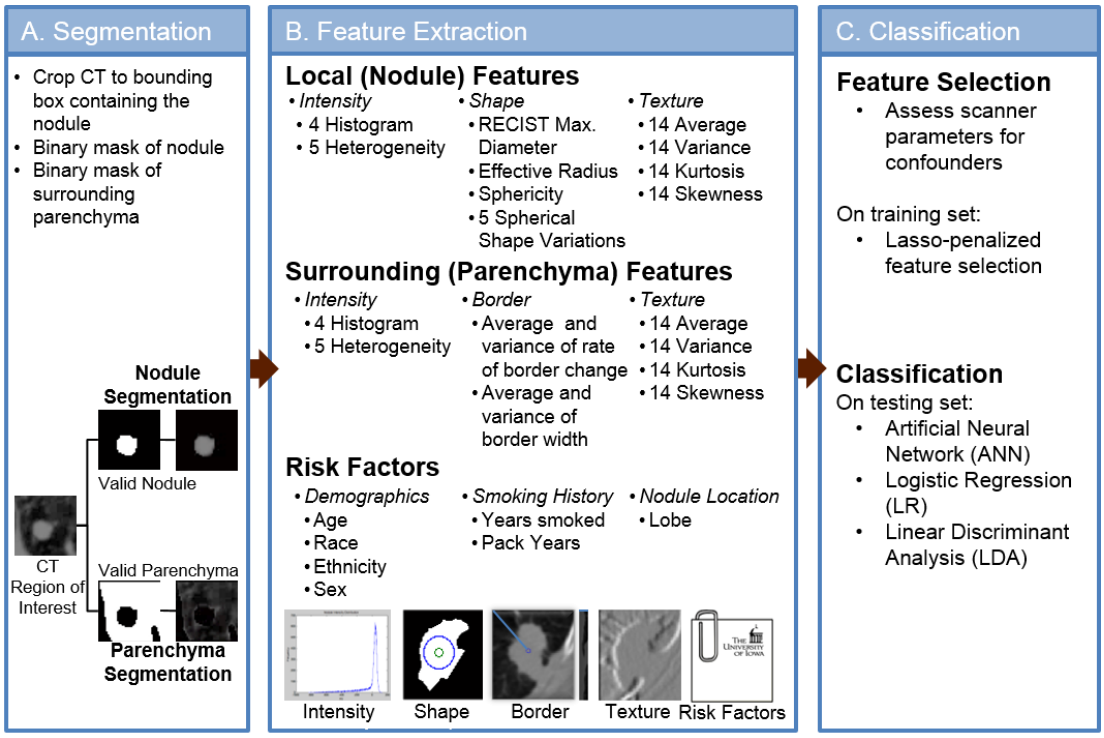

Figure 9. Clinical modifications implemented to the CAD tool include: the use of 2D Laws' TEMs and risk factors in (B) Feature Extraction, a new feature selection method in (C) Classification, as well as the separation of training and testing sets. Finally, three classifiers were explored including ANN, LR, and LDA.

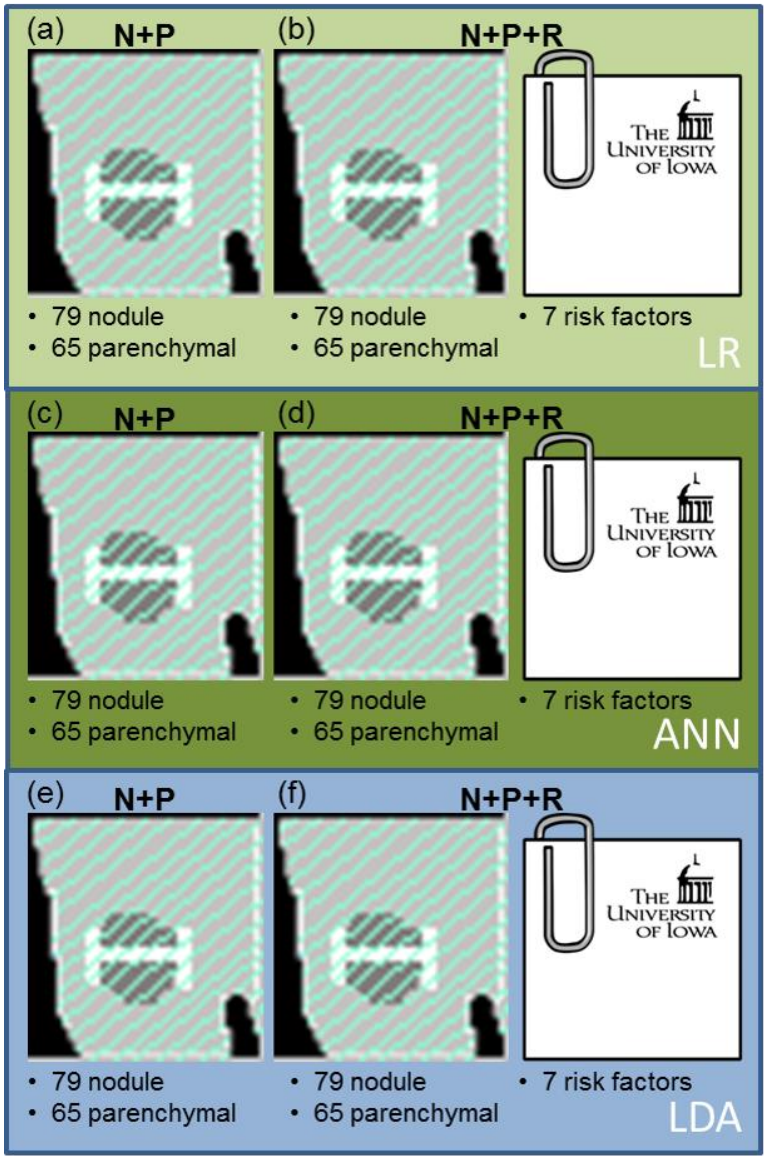

Figure 10. A total of six CAD models were developed. (a) and (b) illustrate the use of logistic regression (LR) for trials built using nodule and parenchyma features alone (N+P) and nodule, parenchyma, and risk factors (N+P+R). (c) and (d) show the N+P and N+P+R trials using an artificial neural network (ANN). Finally, (e) and (f) show the N+P and N+P+R trials constructed with linear discriminant analysis (LDA).

Table 5. Features selected by the least absolute shrinkage and selection operator (lasso) feature selection method. The nodule texture and parenchyma and nodule intensity are important features selected by lasso.

Selected Features	QCT Features (N+P)	QCT Features + Risk Factors (N+P+R)
Lasso	**Intensity** Nodule entropy Parenchyma mean intensity Parenchyma minimum intensity **Shape** Absolute difference mean in boundary Absolute difference variance in boundary **Size** Effective radius **Texture** Nodule skewness 2D texture	**Intensity** Nodule full-width at half maximum Nodule entropy Parenchyma mean intensity **Shape** Absolute difference range in boundary **Texture** Nodule variance 2D texture Nodule skewness 2D texture **Risk Factors** Lobe location Age Sex Pack-years

Table 6. Summary of performance measures for the six CAD models of the clinical cohort. Three classifiers were explored: logistic regression (LR), artificial neural network (ANN), and linear discriminant analysis (LDA). Due to random initialization of the neural networks during training, performance is reported as the mean ± standard deviation over the 10 trials. The area under the ROC curve (AUC) for the training set is reported, as well as the AUC, accuracy, sensitivity, specificity, positive predictive value (PPV) , and negative predictive value (NPV) for the testing set.

Classifier	Clinical Cohort Classifier Performance						
	Training (n=133)	Testing (n=66)					
	AUC	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
QCT Features (N+P)							
LR	0.79	0.68	0.65	0.76	0.52	0.67	0.63
ANN	0.84±0.02	0.66±0.03	0.66±0.06	0.79±0.10	0.49±0.04	0.67	0.65
LDA	0.74	0.68	0.61	0.78	0.38	0.62	0.58
QCT Features + Risk Factors (N+P+R)							
LR	0.93	0.81	0.71	0.73	0.69	0.75	0.67
ANN	0.97±0.01	0.76±0.04	0.70±0.04	0.71±0.05	0.68±0.06	0.74	0.65
LDA	0.88	0.76	0.70	0.70	0.69	0.74	0.65

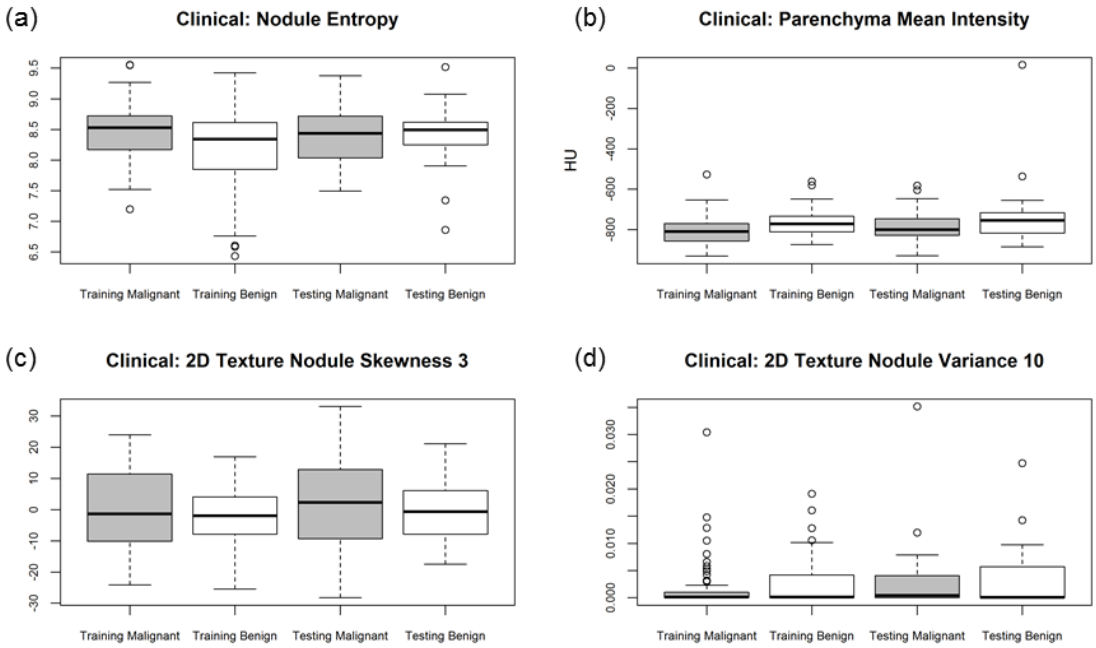

Figure 11. Box plots for clinical features common between the nodule and parenchyma (N+P) and the nodule, parenchyma, and risk factor (N+P+R) CAD trials. Features include: (a) nodule entropy, (b) parenchyma mean intensity, and (c-d) two 2D nodule skewness texture features.

4.3 Discussion

From the feature selection process in the research cohort, 5 of the 53 significant features ($p<0.05$) were parenchymal features, while another 26 were near significance ($p<0.1$). These observations support the hypothesis that inclusion of features from the surrounding parenchyma can aid in the classification of malignant and benign nodules. In the clinical cohort several parenchymal features, including parenchyma mean and minimum intensity, were selected by the N+P and N+P+R classifiers.

Comparing the research and clinical CAD tools, there are several similarities in the features selected. Both CAD tools utilized nodule textures; while the research CAD incorporated 3D textures, the clinical CAD benefited from the modification to 2D textures. Additional work comparing the effects of utilizing 2D vs. 3D trials was conducted with the clinical cohort (Table A 7 and Table A 8). The performances gained by utilizing 3D Laws' TEM features were small for N+P (3D testing AUC: 0.69 vs. 2D testing AUC: 0.68). For N+P+R, the performance with 2D and 3D textures were equivalent (2D and 3D testing AUC: 0.81); however, the CAD utilizing 3D textures relied on a size feature. The benefits of the 2D textures extend beyond equal performances. The 2D Laws' TEMs require less computation time than the 3D Laws' TEMs, accelerating the processing time of each nodule through the CAD tool.

The parenchymal features' influence on classification is also important in regards to the size of lung tumors for which the CAD tool was developed. The CAD tool was developed to aid in the classification of nodules (maximum diameter of 30mm). For small pulmonary nodules, especially those <10mm, the number of voxels from which nodule features can be computed through CT data is limited. By showing the parenchymal features contribute to diagnosis, the number of voxels that can be used in feature extraction is increased. Additionally, inclusion of parenchymal features quantifies the reaction of the nodule to its surroundings. The significance of these features further shows that by examining the nodule alone, valuable classification data is being ignored.

While a statistically significant size bias exists between the malignant and benign nodules in both cohorts (*3.3 Results*), neither of the size features (RECIST maximum diameter and effective radius) were selected for classification in either the research CAD or the clinical N+P+R CAD, suggesting there are stronger predictors of malignancy present in QCT features beyond the inherent size differences. This is important as the majority of nodules identified in CT screening programs will be small after the first CT scan. This first CT screening scan, also called the prevalence time point, identifies nodules which have been developing and growing for some time as well as newly formed nodules. In the research cohort, the majority (38/50) of nodules were detected in this prevalence time point, which accounts for the greater differences in size between the malignant nodules and the benign nodules, which tend to have slower rates of growth. In subsequent scans, also called incidence scans, newly formed nodules will be the primary target; these nodules are not as prone to the differences in size due to the smaller amount of time for growth. The lack of size features selected by the CAD tool for classification indicates a robustness regardless of whether a prevalence or incidence time point is used. Additionally, the most valuable CAD tool will facilitate early diagnosis of CT-identified nodules without requiring longitudinal follow-up imaging and growth pattern data.

While the research cohort presented a moderate amount of variability in terms of dose and manufacturer, the clinical cohort highlights the extensiveness of the feature extraction module. Despite the challenges presented by this clinical cohort, including diverse scan parameters, predominately contrast enhanced scans, and greater variability in demographics and risk factors, the clinical CAD tool was able to identify several robust QCT features, including parenchyma intensity measures and 2D Laws' TEMs. Classifier performance was improved with incorporation of risk factors; the inclusion of risk factors in the decision-making process would serve as a better mimic for a second reader as the physicians also take these factors into account.

Both CAD tools developed in this section benefited from the inclusion of parenchymal features. For the research cohort, the best classifier was the ANN constructed with nodule and parenchyma features (N+P). The clinical CAD tool that yielded best results was the LR classifier that incorporated risk factors as well as nodule and parenchyma features (N+P+R).

CHAPTER V

PARENCHYMAL RING EXPLORATION

We have developed and tested CAD tools that utilize QCT features derived from the surrounding lung parenchyma for both our research and clinical cohorts, completing Aim 1 and demonstrating added benefit with the inclusion of parenchymal features. For Aim 2, we sought to determine the optimal amount of lung parenchyma surrounding a nodule to maximize the CAD tools' performance.

5.1 Rationale

The results of Aim 1 support our hypothesis that the parenchyma provides meaningful information in the classification of lung nodules. In order to further explore the impact of including the parenchyma in the CAD tool, a systematic approach for isolating and quantifying varying amounts of parenchyma around a nodule was required. This section describes this approach and how it was applied in both the research and clinical cohorts.

5.2 Research Cohort

As with the development of the CAD tool, the research cohort was used for the development of our systematic approach to determine the optimal amount of parenchyma to use.

5.2.1 Methods

An additional module was developed and added to the CAD tool pipeline. This preprocessing module takes in the valid parenchyma masks (based roughly on the size of the nodule) and crops them into the desired amount of parenchyma to be used by the *4.1.1.1 Feature Extraction* module.

5.2.1.1 Systematic Parenchyma Selection Approach

A systematic approach was developed, as shown in Figure 12, to generate parenchyma masks that are proportionally related to the size of the nodule. To do this, the maximum nodule diameter was determined from the three-dimensional nodule mask. Next, percentages of the nodule diameter were computed, consisting of 0%, 10%, 20%, 30%, 40%, and 50% of the length of the diameter, in millimeters. Voxels within these distances from the border of the nodule were selected to be part of the parenchymal masks. Two dimensional slices of these differing parenchyma masks can be seen in Figure 13, including the composite image of the six parenchyma masks overlapped in order to compare the amount of parenchyma used for each mask.

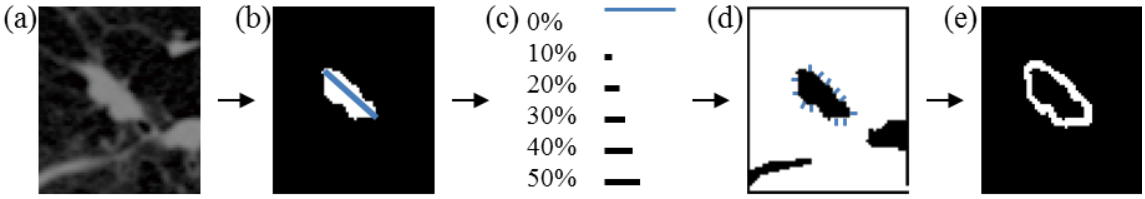

Figure 12. Parenchyma Mask Development. From the region of interest, two-dimensional slice shown in (a), the nodule region is segmented, shown in white (b). The maximum diameter for the nodule is determined, shown by the blue line in (b), and percentages of the diameter are computed (c). In (d), the valid parenchyma voxels, in white, that are less than 10% length of the diameter from the nodule border, indicated by blue lines, are included in the final parenchyma mask, (e). (d) and (e) are then repeated for 0%, 20%, 30%, 40%, and 50% of the length of the diameter, generating 6 parenchyma masks.

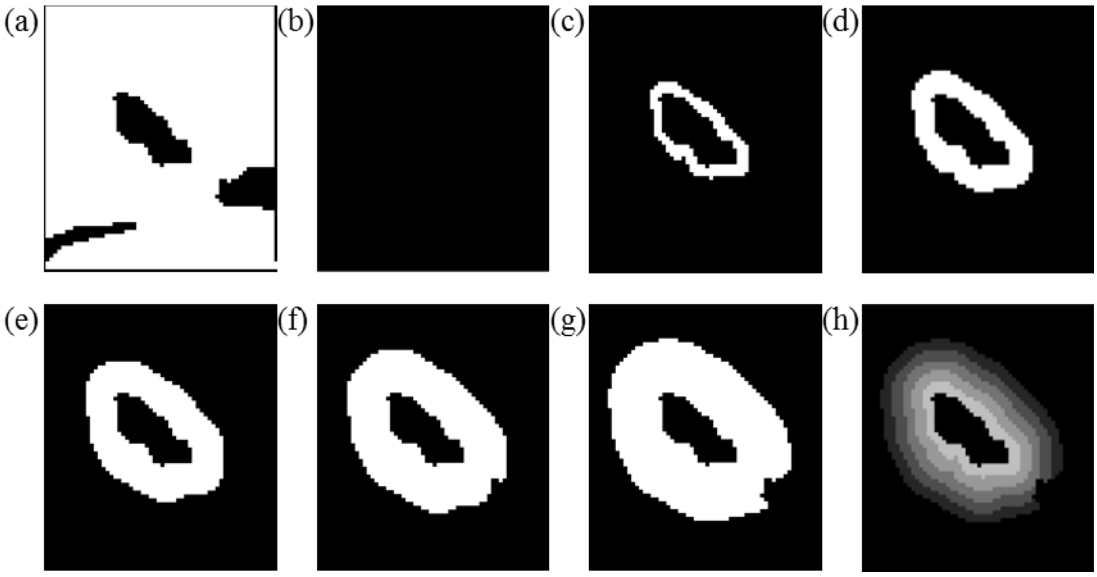

Figure 13. A two-dimensional slice of a region of interest, showing the varying amounts of parenchyma used. (a) shows the total valid parenchyma of the region of interest. (b), (c), and (d) on the top row show the parenchyma within 0%, 10%, and 20% of the maximum diameter length from the nodule's border, respectively. Similarly, (e), (f), and (g) show the voxels within 30%, 40%, and 50% of the maximum diameter length from the nodule's border. (h) shows the six parenchyma masks layered to compare the different parenchyma masks.

5.2.1.2 Optimal Ring Identification

Upon creation of the six parenchyma masks, the 50 research nodules were processed through the CAD tool optimized in *4.1.1.2 Feature Selection and Classification*. Each parenchyma mask served as a separate trial, resulting in six classifiers. The number of significant features identified through statistical analysis, the features selected for classification, and the performance of the CAD tool were examined and compared between the six classifiers. The five-case testing set was processed through the six classifiers to ensure overfitting had not occurred.

5.2.1.3 Isolating the Parenchymal Signal Source

Upon identification of the optimal amount of parenchyma for inclusion in the CAD tool, the parenchyma mask selected was explored for the source of the signal. In order to discover whether it is the volume of parenchyma included or the volume as well as its proximity to the nodule that is most important in improving CAD performance, an additional parenchyma mask, the peel mask, was constructed. The inner mask, the 10% parenchyma mask, was subtracted from the outer parenchymal mask one size larger (limited by the 50% mask) than the optimal parenchymal mask identified in *5.2.1.2 Optimal Ring Identification*. This process leads to the generation of a peel mask, designed to remove the local effects of the nodule in order to assess the importance of proximity of the parenchyma to the nodule. An example of this process is shown in Figure 14, where 30% was the optimal ring, resulting in 40% being selected as the outer parenchymal mask.

Figure 14. Construction of peel mask. (a) the larger parenchyma mask (in white), in this illustration the 40% mask, is selected for the outer bounds of the parenchyma mask. The parenchyma mask to be removed (b, also in white), the 10% mask, is subtracted from (a), leaving the outer peel of parenchyma (c). The nodule volume is shown in grey in (a)-(c).

Parenchyma features from this peel mask were extracted, along with nodule features, and the nodule and peel features were processed through feature selection and

the ANN classifier. Performance was assessed through accuracy, sensitivity, specificity, and AUC, as well as performance on the blinded testing set.

5.2.1.4 Global Indicators of Disease

In addition to determining the source of the signal in the parenchyma mask, global features were included to determine whether imaging data captured within the same lobe and/or lung contain meaningful information related to nodule diagnosis. Measures including lobar and lung mean and standard deviation intensity, tissue volume, air volume, and emphysema measures such as percent emphysema (percent of voxels < -950 HU on inspiratory imaging) and percent gas trapping (percent of voxels < -856 HU on expiratory imaging) were extracted through Apollo Software (VIDA Diagnostics, Inc., Coralville, IA). A complete list of global features can be found in Table A 6.

Upon identification of the optimal parenchyma mask, these 54 global features were included in the CAD tool for feature selection and classification (see *4.1.1.2 Feature Selection and Classification*). Assessment of CAD performance measures included AUC, accuracy, sensitivity, specificity, PPV, and NPV and comparisons to the N, N+P, and N+Optimal Percent CAD tool. Classifier ROC curves were compared using the Delong method to assess for differences in performances [46]. Predicted classification probabilities were averaged over the 10 trials for comparisons of ANN performance. All statistical comparisons were two-sided and assessed for significance at the 5% level.

5.2.2 Results

The 55 research cases were processed through the parenchyma preprocessing step to generate the six parenchyma masks based on nodule size. Using the 50-case training set, each set of parenchyma masks were processed through the CAD tool, resulting in 6 classifiers. The 10% mask trial had 5 cases whose parenchyma masks were empty due to 10% of the nodule diameter being sub-voxel. These cases were excluded from feature selection and classification, thus this cohort is not directly comparable to the other

percentage masks. The 5-case testing set was used to verify the classifiers did not suffer from over-fitting.

5.2.2.1 Optimal Ring Identification

The number of significant features and performance of the six classifiers were compared to identify the optimal amount of parenchyma to use for nodule classification. As seen in Figure 15, as the amount of parenchyma increased, the number of statistically significant features increased until the amount of parenchyma exceeded 30%. At this point, the number of parenchymal features decreased.

Upon completion of the stepwise forward selection with majority voting, the most commonly selected features for the six classifiers included: parenchyma variance, parenchyma entropy, nodule variance, absolute difference in range between the nodule boundary and the equivalent volume sphere, and the standard deviation of the border width. The 10% mask selected 3 features for classification before the accuracy did not improve with additional features. All other classifiers selected five features.

Classification performances of the CAD tools ranged from 89.8% (44.9/50) to 93.2% (46.6/50), with the maximum classifier performance occurring for the 30% mask with the size feature of effective radius (Table 7). Excluding size as a feature, the best performing mask was the 50% mask with an accuracy of 92% (46/50); however, the 20% and 40% masks had similar classifier performances of 91.0-91.6% (45.5-45.8/50) and AUCs of 0.86. Additionally, these classifiers do not greatly vary from the N+P classifier (accuracy of 91.6% (45.8/50)). While the performance is not improved, the use of a systematic approach for exploring the parenchyma is preferred in order to minimize the amount of variability introduced by subjective decisions.

Looking at the overall trends (Figure 16), there is a global trend for the performance of the classifiers to stabilize after 3-4 features are included in classification; the addition of more features does not improve accuracy. However, the 30% mask with

size drastically differs, with the classifier performance continuing to increase as more features are included in analysis; additionally, the 50% mask does not suffer from this effect as greatly as the other affected masks.

Improved performance in these masks is also seen in the blind testing set. While the 30% with size and 50% masks had 100% accuracy on the test cases, the 20% had an accuracy of 80% (1/2 malignant and 3/3 benign) and the 40% mask had an average accuracy of 94% (1.68/2 malignant and 3.3 benign). The 10% performance was challenged due to two benign nodules having empty parenchyma masks; therefore, the accuracy was 60% (2/2 malignant and 1/3 benign).

While the increased number of significant features, the use of parenchyma features in classification, and the steady increase in classifier performance of the 30% mask shows promise, its use of size as a feature in classification introduces limitations in the CAD's applicability to smaller, incidence nodules in lung cancer screening. Therefore, the 50% parenchyma mask was selected as the ideal percent of parenchyma to include due to its systematic approach of isolating parenchyma from the valid parenchyma mask and its slightly improved performance over the other percentages.

5.2.2.2 Isolating the Parenchymal Signal Source

Upon the selection of the 50% mask as the optimal percentage of parenchyma to include, the process outlined in *5.2.1.3 Isolating the Parenchymal Signal Source* was performed using the 50% mask as the outer mask.

Compared to the 50% mask, which had 78 significant features, the (50-10)% peel mask had 95 significant features including 6 nodule intensity, 5 parenchymal intensity, 2 shape, 2 size, 3 border, 34 nodule texture, and 43 parenchymal texture features. Despite the increase in significant features, the features appear to be weaker separators of malignancy. Feature selection chose 2 parenchyma skewness textures, 1 shape feature, 1 parenchyma intensity feature, and 1 nodule intensity feature. The AUC for the

N+(50-10)% trial was 0.92, an improvement compared to the AUC of 0.86 for the N+50% trial; however, the accuracy of this classifier was 90.1%, reduced from the accuracy of the N+50% trial at 92.0%. Additionally, poor performance on the independent cohort (accuracy of 24% (0/2 malignant and 1.2/3 benign)) suggests overfitting occurred.

5.2.2.3 Global Indicators of Disease

The nodule and parenchyma features extracted using the N+50% CAD were coupled with the global features, resulting in a N+50%+G ANN trial. The features selected are shown in Table 8 and box plots are shown in Figure 17. Additional comparisons of features selected can be seen in Table A 10. The parenchyma entropy and two features from the lobe containing the nodule were selected for classification. The parenchyma entropy was lower in the benign nodules, indicating less randomness within the parenchyma of these nodules. Similarly, the intensity of the lobe containing the nodule had a lower standard deviation for benign nodules.

The N+50%+G CAD tool performed very well, with an AUC of 0.94 (Figure 18). The N+50%+G CAD tool performed better compared to the N and N+P CAD tools developed in 4.1 Research Cohort, with AUCs of 0.88 and 0.91, respectively. While not a statistically significant improvement (smallest $p=0.17$ for N vs. N+50%+G), the modest sample size limits the power to detect small differences. Results of blinded testing showed modest performance of 60% accuracy (1/2 malignant and 2/3 benign). The optimized N+50%+G CAD tool shows a trend toward increased performance in all performance metrics.

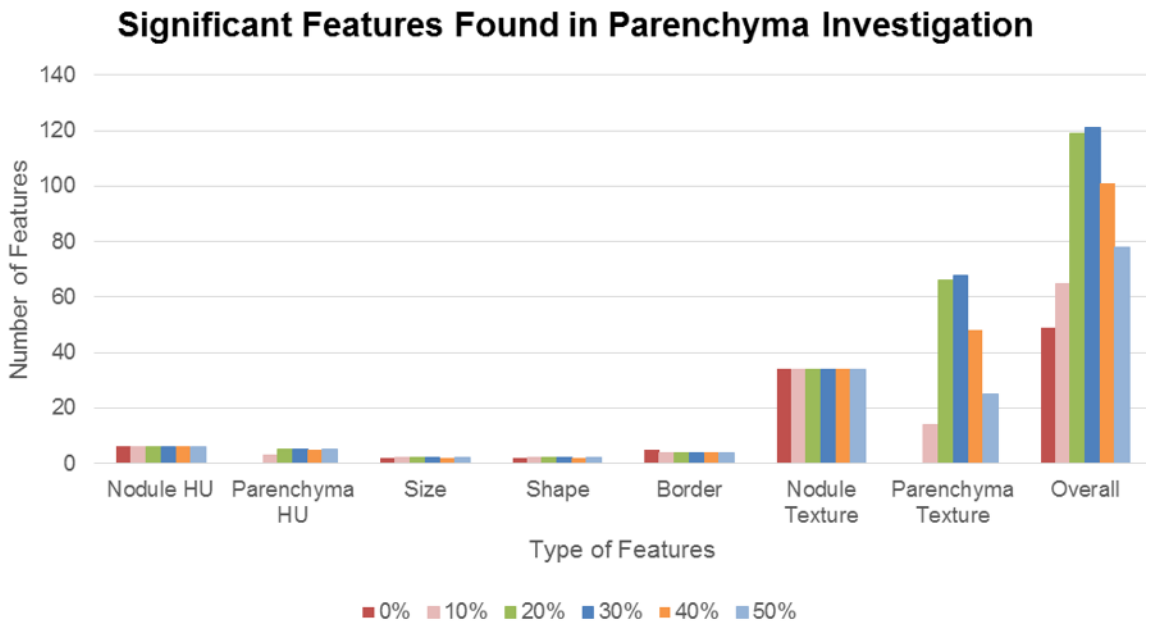

Figure 15. Statistical features from parenchymal investigation. The number of significant features increases as the amount of parenchyma increases, with the maximum number of significant features identified when the 30% mask is used.

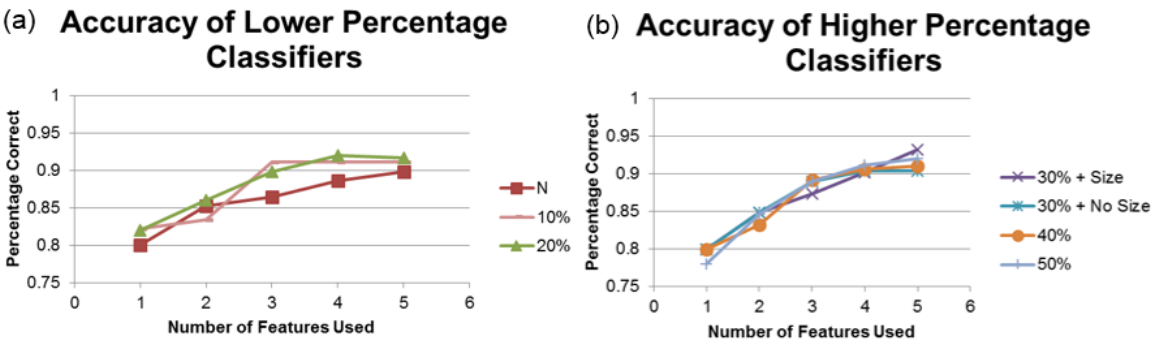

Figure 16. Classifier performances as the number of features included is increased. In (a), the accuracies of the lower percentage classifiers tend to level off after 3 features are used for classification. However, for the higher percentage classifiers (b), the only percentage that continues to increase with more features is the 30%+size mask. The 40% and 50% masks' performances level off after 4 features, respectively.

Table 7. Summary of performance measures, averaged over ten trials, for the ANN classification of the research cohort using varying amounts of parenchyma. The positive predictive value (PPV), negative predictive value (NPV), and the area under the ROC curve (AUC) are also reported. Note: N+10% is out of 45 cases.

Selected Features	ANN Leave-one-case-out Classifier Performance						
	# correct 50 (18, 32)	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Nodule Features: N Nodule Variance Nodule Kurtosis TEM17 Mean of Border Nodule Kurtosis TEM2 Nodule Kurtosis TEM16	44.9 (15,29.9)	89.8±0.6%	83.3±0.0%	93.4±1.0%	.877	.908	0.882±0.005
Nodule & Parenchyma Features: N+10% Nodule Variance Par. Kurtosis TEM3 Par. Kurtosis TEM2	41.0 (14,27)	91.1±0.0%	82.4±0.0%	96.4±0.0%	.933	.900	0.821±0.006
Nodule & Parenchyma Features: N+20% Parenchyma Entropy Parenchyma Variance Parenchyma Mean HU Mean of Columns Par. Skewness TEM33	45.8 (13.9,31.9)	91.6±0.8%	77.2±1.8%	99.7±1.0%	.993	.886	0.867±0.013
Nodule & Parenchyma Features: N+30% Parenchyma Entropy Parenchyma Variance Nodule Kurtosis Effective Radius St. Dev. Of Border	46.6 (14.8,31.8)	93.2±1.0%	82.2±2.3%	99.4±1.3%	.967	.909	0.865±0.014
Nodule & Parenchyma Features: N+30% -S Parenchyma Entropy Parenchyma Variance Nodule Kurtosis Abs. Difference Range St. Dev. Of Border	45.2 (14.3,30.9)	90.4±2.3%	79.4±2.7%	96.6±2.7%	.929	.893	0.866±0.012
Nodule & Parenchyma Features: N+40% Parenchyma Entropy Nodule Median HU Abs. Difference Range Parenchyma Variance St. Dev. Of Border	45.4 (14,31.5)	91.0±1.4%	77.8±0.0%	98.4±2.2%	.966	.887	0.869±0.034
Nodule & Parenchyma Features: N+50% Parenchyma Entropy Nodule Median HU Abs. Difference Range Parenchyma Variance Nodule Variance	46.0 (14,32)	92.0±0.0%	77.8±0.0	100±0.0%	1.00	.889	0.867±0.027

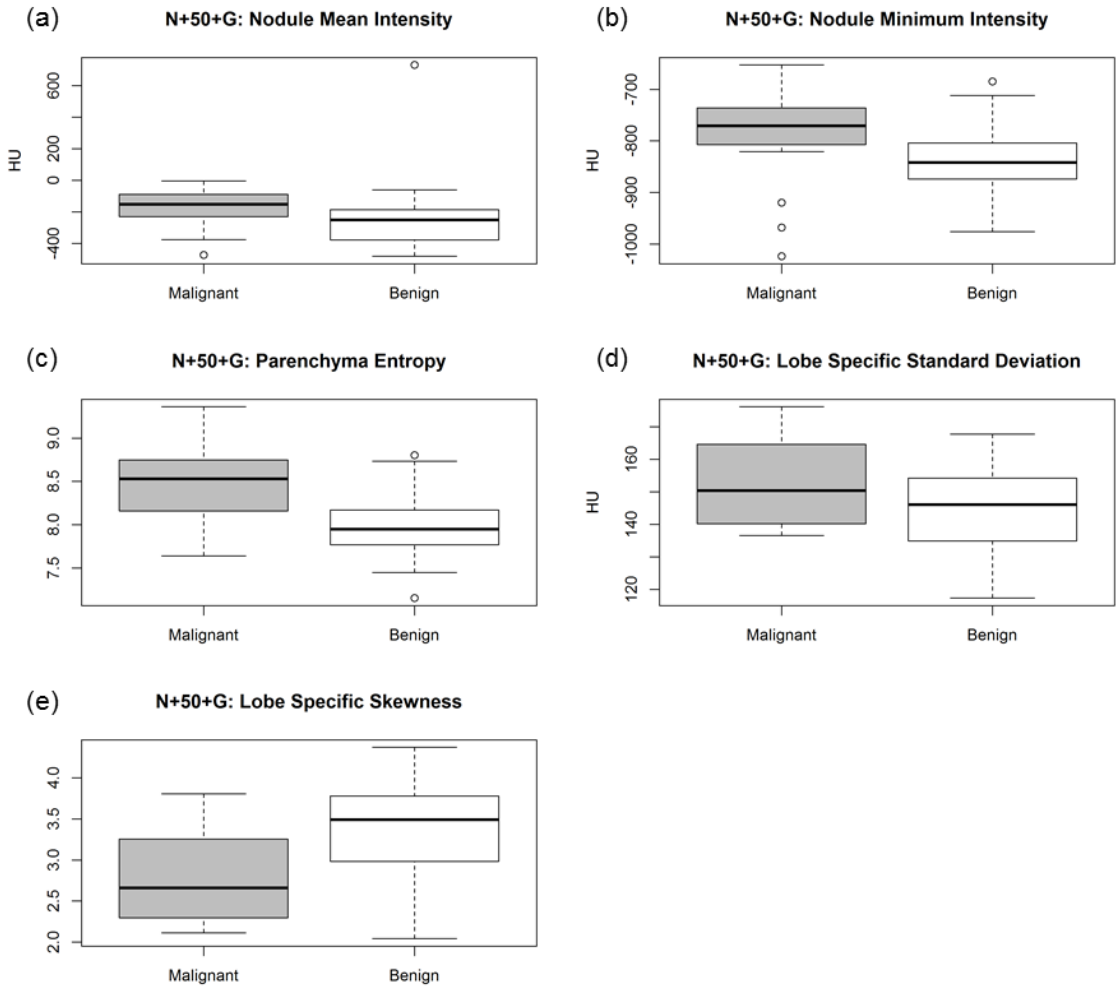

Figure 17. Boxplot for N+50%+G features selected during feature selection. Malignant ROIs have greater entropy (c) in the surrounding parenchyma as well as more variance in the lobe (d) compared to benign nodules.

Table 8. Summary of performance measures, averaged over ten trials, for the ANN classification of the research cohort using N, N+P, N+50%, N+(50-10)%, and N+50%+G features. The positive predictive value (PPV), negative predictive value (NPV), and the area under the ROC curve (AUC) are also reported.

Selected Features	ANN Leave-one-case-out Classifier Performance						
	# correct 50 (18, 32)	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Nodule Features: N Nodule Variance Nodule Kurtosis TEM17 Mean of Border Nodule Kurtosis TEM2 Nodule Kurtosis TEM16	44.9 (15,29.9)	89.8±0.6%	83.3±0.0%	93.4±1.0%	.877	.908	0.882±0.005
Nodule & Parenchyma Features: N+P Parenchyma Variance Nodule Minimum HU Nodule Kurtosis TEM1 Nodule Kurtosis TEM15	45.8 (16,29.8)	91.6±0.8%	88.9±0.0%	93.1±1.3%	.879	.937	0.913±0.019
Nodule & Parenchyma Features: N+50% Parenchyma Entropy Nodule Median HU Abs. Difference Range Parenchyma Variance Nodule Variance	46.0 (14,32)	92.0±0.0%	77.8±0.0%	100±0.0%	1.00	.889	0.867±0.027
Nodule & Parenchyma Features: N+(50-10)% Par. Skewness TEM29 Abs. Difference Range Par. Skewness TEM34 Nodule Variance Parenchyma Entropy	45.4 (15.2, 30.2)	90.8±1.7%	84.4±2.3%	94.4±2.0%	.894	.915	0.923±0.021
Nodule , Parenchyma & Global Features: N+50%+G Parenchyma Entropy Lobe-specific Skewness Nodule Minimum HU Nodule Mean HU Lobe-specific standard deviation	47.8 (16.6,31.2)	95.6±1.6%	92.2±2.9%	97.5±2.5%	.954	.957	0.941±0.000

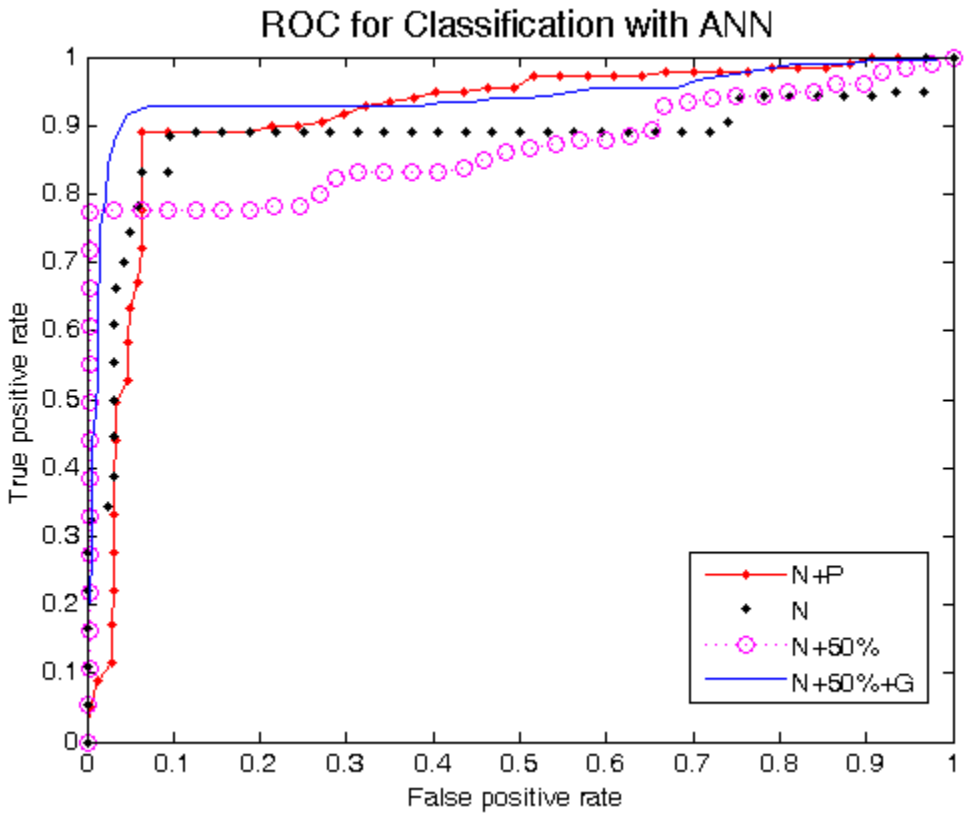

Figure 18. ROC curves for N, N+P, N+50%, and N+50%+G CAD tools. The N+50%+G ROC curve (blue) shows better performance than the N (black), N+P (red), or N+50% (purple) ROC curves.

5.3 Clinical Cohort

Based on the observations from the research cohort, the framework for identifying the optimal amount of parenchyma to use in the clinical cohort was developed. The CAD tools were trained with the 133-case training set and the performances assessed using the blinded 66-case testing set, as described in 3.3.2 Clinical Cohort.

5.3.1 Methods

Using the procedure described in *5.2.1.2 Optimal Ring Identification*, the clinical cases' valid parenchyma masks were sent through the preprocessing module, resulting in each nodule having 6 parenchyma masks. As the research cohort illustrated the 30% and 50% mask contain the most promise, the other nodule masks were left out of analysis. For the 10% mask, this was out of practicality. In the research cohort, five nodules were excluded due to 10% of their diameter being sub-voxel. In the clinical cohort, where the voxels are larger, this was a bigger issue. The nodule only (0%) mask was included for comparison to a nodule-only clinical CAD tool.

The 0%, 30%, and 50% masks were processed through the optimized clinical CAD tool (*4.2.3.2 Classification*), resulting in 3 classifiers. The features selected and the performances of these classifiers were assessed to determine the optimal amount of parenchyma to include in the final clinical CAD tool.

5.3.2 Results

Using the 133-case training set and the 66-case testing set, 0%, 30% and 50% masks were processed through the N+P+R LR clinical CAD tool. The features selected through lasso and the performances of these CAD tools can be seen in Table 9. Similar to the research CAD tool, while the 50% mask modestly outperformed the other trials, the use of a systematic approach to determine the amount of parenchyma to include is an important procedure. Further, CAD performance using any amount of parenchyma is improved over the nodule only (N+0%+R) CAD tool. Box plots showing the trend in

features selected by the 50% mask are shown in Figure 19. The means, standard deviations, and p-values for the quantitative features selected by this CAD tool can be found in Table A 11.

Table 9. Summary of performance measures for the clinical cohort using varying amounts of parenchyma. The N+P+R features selected by lasso-penalized logistic regression are shown, and the area under the ROC curve (AUC) for the training and testing set are also reported.

Amount of Parenchyma	Clinical N+P+R LR Classifier		
	QCT Features + Risk Factors (N+P+R)	Training AUC	Testing AUC
0%	**Intensity** Nodule variance Nodule full-width at half maximum Nodule entropy **Shape** Absolute difference range in boundary **Border** Standard deviation of slope **Texture** Nodule variance 2D texture Nodule skewness 2D texture **Risk Factors** Lobe location Age Sex Pack-years	0.94	0.77
30%	**Intensity** Nodule variance Nodule full-width at half maximum Nodule entropy Parenchyma median intensity **Shape** Absolute difference range in boundary **Texture** Nodule variance 2D texture Nodule skewness 2D texture Parenchyma variance 2D texture **Risk Factors** Lobe location Age Sex Pack-years	0.94	0.79
50%	**Intensity** Nodule variance Nodule full-width at half maximum Nodule entropy Parenchyma median intensity **Shape** Absolute difference range in boundary **Texture** Nodule variance 2D texture Nodule skewness 2D texture Parenchyma variance 2D texture **Risk Factors** Lobe location Age Sex Pack-years	0.94	0.80

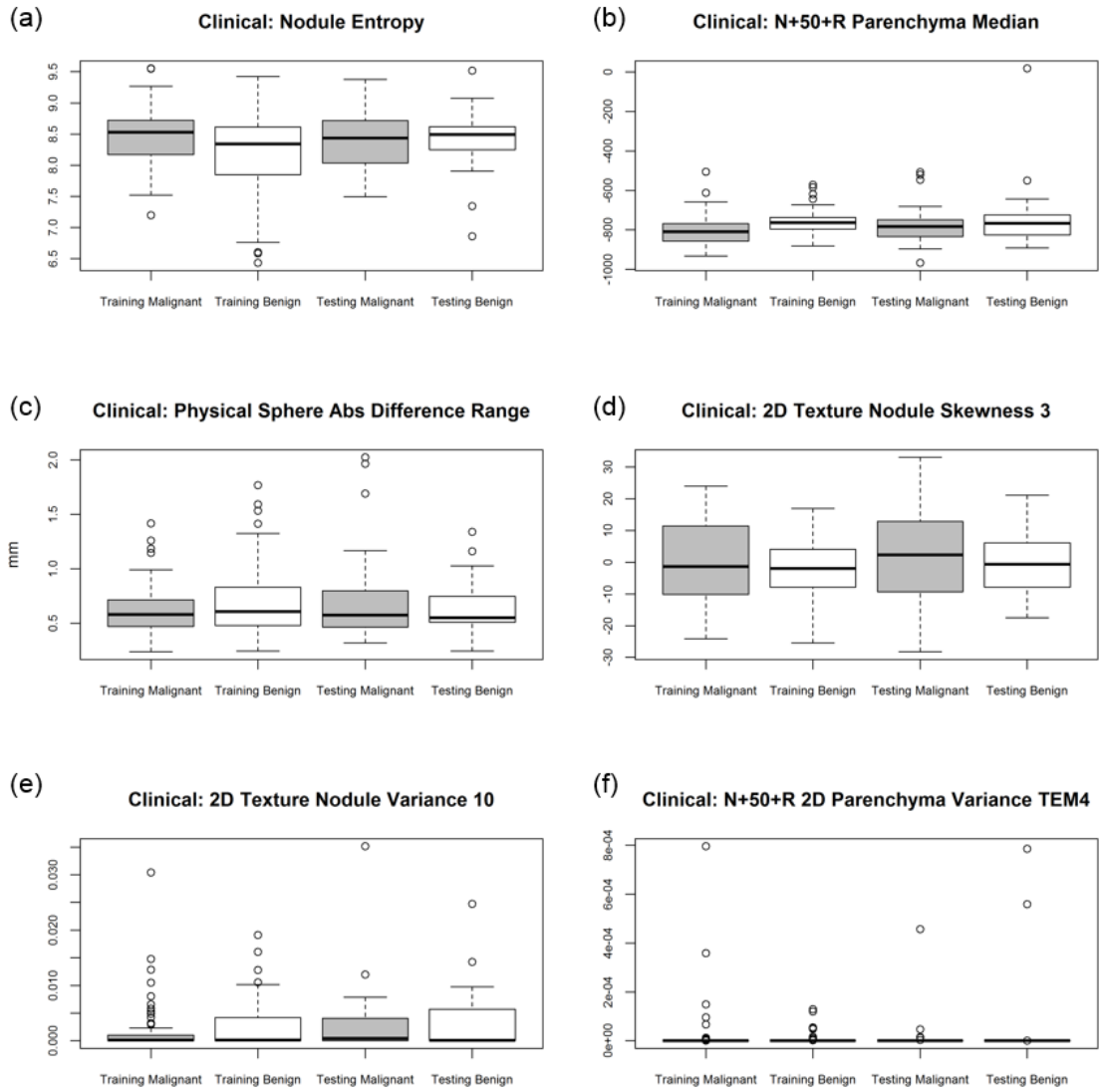

Figure 19. Boxplots for N+50%+R features selected during lasso feature selection. Nodule entropy (a) and texture nodule skewness (d) are higher for malignant nodules than benign nodules, whereas the benign nodules have larger texture nodule variance (e) and parenchyma median intensity (b).

5.4 Discussion

For the research cohort, the optimized research-driven CAD tool is the N+50%+G ANN, constructed from nodule features, parenchyma features from the 50% mask, and global features from the lobe containing the nodule. Similarly, the optimal clinically focused, lung cancer screening CAD tool is the N+50%+R LR model. This CAD tool, constructed through lasso-penalized logistic regression, relies on features from the nodule, 50% parenchyma mask, and clinical risk factors.

As we saw in the development of the CAD tools, the presence of size bias had an impact on this study. The research 30% mask performed very well, but relied heavily on size. When size was removed as a feature, the performance of the CAD tool suffered. However, the 50% mask performed strongly without size as feature, and, when global features were included, improved its performance further in nodule classification. The addition of more parenchyma illustrates a clear trend to overcoming size bias and providing useful information to the CAD tool.

Querying the parenchyma rings for the source of the signal yielded interesting results. While the (50-10)% peel mask resulted in an increase in statistically significant features compared to the 50% parenchyma mask, the accuracy and generalizability during classification was lower, suggesting that, while more features are significant, the 50% mask contains stronger sources of signal. Hence, the interaction between the nodule and parenchyma is important.

These results also demonstrate the benefits of using a systematic approach to determine the amount of parenchyma to include. By relating the amount of surrounding parenchyma directly to the size of the nodule, the number of significant parenchyma features increased compared to the number of significant features used with all the valid parenchyma within the ROI.

In addition to explore the source of signal in the research CAD tool, we sought to determine whether the scan variability was influencing the performance of the clinical

CAD tool. We removed the 15 non-contrast CT scans from the training and testing sets, resulting in 123 contrast-enhanced training cases and 60 testing cases. The N+50%+R CAD tool selected similar features and had equivalent performance (training AUC: 0.95, testing AUC: 0.79, see Table A 9). This provides strong evidence that our CAD tool is selecting features robust to variations in protocol.

For Aim 2, we sought to determine the impact of including parenchymal features by optimizing the amount of parenchyma surrounding a nodule to include in the CAD tools. For the research-driven CAD tool, the N+50%+G ANN was found to be the optimal CAD tool. Not only does this CAD tool utilize a systematic approach for the inclusion of a standard amount of parenchyma, lobar measures of lung disease are also considered. For the clinical CAD tool, the 50% mask also rose as the ideal amount of parenchyma with the N+50%+R LR being selected as the optimal CAD tool.

CHAPTER VI

IMPLEMENTING A THREE-CLASS APPROACH

For Aim 3, we sought to explore the preliminary clinical impact of a CAD tool that includes parenchymal features. One way of further investigating the clinical impact of such a CAD tool was the development and implementation of a three-class approach.

6.1 Rationale

According to the National Comprehensive Cancer Network Guidelines, solid nodules greater than 6mm require either: multiple repeated CT imaging, PET/CT imaging, or an invasive procedure to determine presence of malignancy [12]. The long-term motivation for this project is to develop a CAD tool that will reduce the number of follow-up procedures, including repeated imaging studies, invasive biopsies, and resections, than would be required based on CT imaging alone by increasing the confidence of the radiologist in the CAD predicted diagnosis of a nodule. However, even with the high accuracy of the CAD tools we have developed thus far, a small chance of misclassification is likely to undermine the trust of the clinician in the CAD tool. Therefore, in a clinical setting, a three-stage diagnosis decision tool may be of greater value by dividing the cases into: (1) a definite malignant class, where the patient would be sent straight to intervention, (2) a definite benign class, where the patient would continue to receive low-dose annual CT screening, and (3) an indeterminate class who would undergo the standard-of-care follow-up.

6.2 Research Cohort

The ANN classifier used by the optimized research-driven CAD tool outputs a numeric value, which is then subjected to thresholding in order to determine the binary class decision. Taking advantage of these numeric outputs, we explored implementing a double-threshold, three-classification scheme.

6.2.1 Methods

Using the ANN classifier outputs from the optimized research CAD tool (N+50%+G), two thresholds were selected to divide the classification space into malignant, benign, and indeterminate. First, the ANN outputs of 10 trials were averaged. The thresholds were selected based on the averaged ANN output values of the most extreme misclassified malignant (lower threshold) and benign (upper threshold) case. A buffer of one standard deviation of each class was incorporated to allow for more extreme values to be seen. Based on these thresholds, every nodule above the upper threshold (the most extreme benign value + one standard deviation of benign cases) was correctly labeled as malignant, and every nodule below the lower threshold (the most extreme malignant value - one standard deviation of malignant cases) was correctly labeled as benign. Those cases falling in between thresholds received the label of indeterminate. In a clinical setting, the CAD tool would recommend patients with these nodules undergo follow-up using the NCCN Guidelines, while those in the definite malignant category would go directly to treatment and those falling in the definite benign category would continue receiving low-dose annual CT screening. After the thresholds were determined, they were applied in a similar manner to the 5-case blinded testing set to assess the generalizability of such a technique.

6.2.2 Results

The two threshold approach was applied to the optimized research-driven CAD tool, which was constructed with nodule features, parenchyma features from the 50% ring, and global features (N+50%+G) (Figure 20). When a traditional two-class approach was applied to the averaged ANN outputs, the sensitivity of the classifier (the accuracy of the malignant category) was 88.9% (16/18 malignant nodules correctly identified). The specificity was 96.9% (31/32 benign nodules correctly identified). Clinically, this would

have resulted in 3 nodules that were misclassified: 2 malignant nodules that would have been missed and 1 benign nodule that would have undergone further follow-up.

Applying the 2 threshold/3 class approach to the N+50%+G ANN outputs, the thresholds fell at 0 and 0.67 (Figure 20). For the 18 malignant cases, 15 were assigned to the definite malignant category; the remaining 3 malignant cases and all 32 benign cases were labeled as indeterminate. Using this three class approach would have sent 15 patients with malignant nodules straight to treatment, reducing the number of follow-up procedures by $\frac{50-35}{50} \times 100 = 30\%$.

There was one extreme malignant case which resulted in the lower threshold being set to 0. This misclassified malignant case was such that the ANN output value (before thresholding into binary classes) falls near 0, which can be interpreted as a strongly benign classification. This COPDGene participant verbally confirmed they had lung cancer but refused to join the Lung Cancer Registry, removing the ability to confirm malignancy through medical records and tissue diagnosis. This participant is now deceased (20 months after the CT scan utilized in this study was acquired), with cause of death reported as COPD and lung cancer. Without pathology, there is doubt as to whether this CT-certified nodule is the cause of the participant's lung cancer, or if this nodule was a benign process and a second malignant lesion developed after Phase I of COPDGene.

When this case was removed from the threshold placement decision, the lower threshold was raised to 0.11, narrowing the indeterminate band (Figure 21). Again, 15 malignant cases were assigned to the definite malignant category. However, 27 of the 32 benign cases were labeled as definite benign. This trial resulted in 7 nodules falling in the indeterminate category, drastically reducing the number of cases referred to follow-up in clinic by $\frac{49-7}{49} \times 100 = 85.7\%$.

The 5 blinded cases were a challenging test for generalizability on the three-class approach. When binary classification was performed, one malignant and one benign nodule were misclassified. With the three-class approach build using all 50 cases, the

benign nodule was still misclassified, but the previously missed malignant nodule fell in the indeterminate category with the other two benign nodules (Figure 22). However, using the 49-case thresholds (0.11 and 0.67), these two nodules are again misclassified (Figure 23). The performance on the blinded test set highlights the limitations of post-classification thresholding and supports further development of this approach.

6.3 Clinical Cohort

In a similar manner to the research cohort, a two threshold/three class approach was applied to the clinical cohort. However, modifications were made due to the independence of the training and testing sets.

6.3.1 Methods

The predictive outcomes from the optimized clinical CAD tool (the LR-trained N+50%+R) were collected for the training and testing set. Using the training set, the most extreme misclassified malignant and benign nodules were identified. The misclassified malignant nodule was used to establish the lower threshold, while the most extreme misclassified benign nodule was used to determine the upper threshold value. Similarly to the research cohort, a buffer of one standard deviation of variability per class was added to each threshold.

These thresholds were applied to the testing set's predictive outcomes from the optimized clinical CAD tool. The number of nodules correctly falling in the definite malignant and definite benign nodules were reported for the training and testing sets, and the percentage reduction of cases to follow-up was computed. The number of nodules assigned to the indeterminate category was also determined. Of utmost importance, misclassified cases (malignant cases falling in the definite benign and vice versa) for the testing set were identified, as this was strongly undesirable and would result in additional testing when none was needed, or worse, a missed cancer.

6.3.2 Results

The two threshold approach was applied to the optimized clinical CAD tool, which was constructed with nodule features and parenchyma features from the 50% parenchyma mask, including two-dimensional textures and risk factors (N+50%+R) (Figure 24). When a traditional two-class approach was applied to the LR predictive outcome (threshold at 0.5), the sensitivity of the classifier (the accuracy of the malignant category) was 86.1% (62/72 malignant nodules correctly identified) for the training set and 75.6% (28/37) for the testing set. The specificity was 81.9% (50/61 benign nodules correctly identified) for the training set and 69.0% (20/29) for the testing set. Clinically, this would have resulted in 18 nodules from the test set that were misclassified: 9 malignant nodules that would have been missed and 9 benign nodules that would have undergone further follow-up.

Applying the 2 threshold/3 class approach to the training set's N+50%+R LR outputs, the thresholds fell at 0.117 and 0.979 (Figure 24). These thresholds were then applied to the testing set of 66 nodules (Figure 25). For the 37 malignant cases, 3 were assigned to the definite malignant category; 34 malignant nodules were labeled as indeterminate, and no nodules were misclassified as benign. For the 29 benign cases, 5 were assigned to the definite benign category, 24 nodules were labeled as indeterminate, and, again, no benign nodules were misclassified as malignant. Using this three class approach would have sent 3 patients with malignant nodules straight to treatment and 5 patients with benign nodules to annual screening, reducing the number of follow-up procedures by $\frac{66-58}{66} \times 100 = 12.1\%$. While this is a modest reduction, no malignant nodules were missed and no patients would have undergone treatment for benign nodules.

6.4 Discussion

The use of a three-class, two-threshold approach shows promise in its ability to reduce the number of unnecessary imaging and invasive procedures and the risks associated with these. As we see with the clinical cohort, post-classification thresholding using the training data is sensitive enough to achieve perfect classification within the definite malignant and benign categories; however its specificity could be improved to further reduce the number of follow-up procedures. The small research test set, however, lacked this sensitivity, misclassifying one benign and one malignant nodule. Therefore, a classification schema integrating a cost function in order to assign a nodule into one of the three classes would be advantageous to pursue.

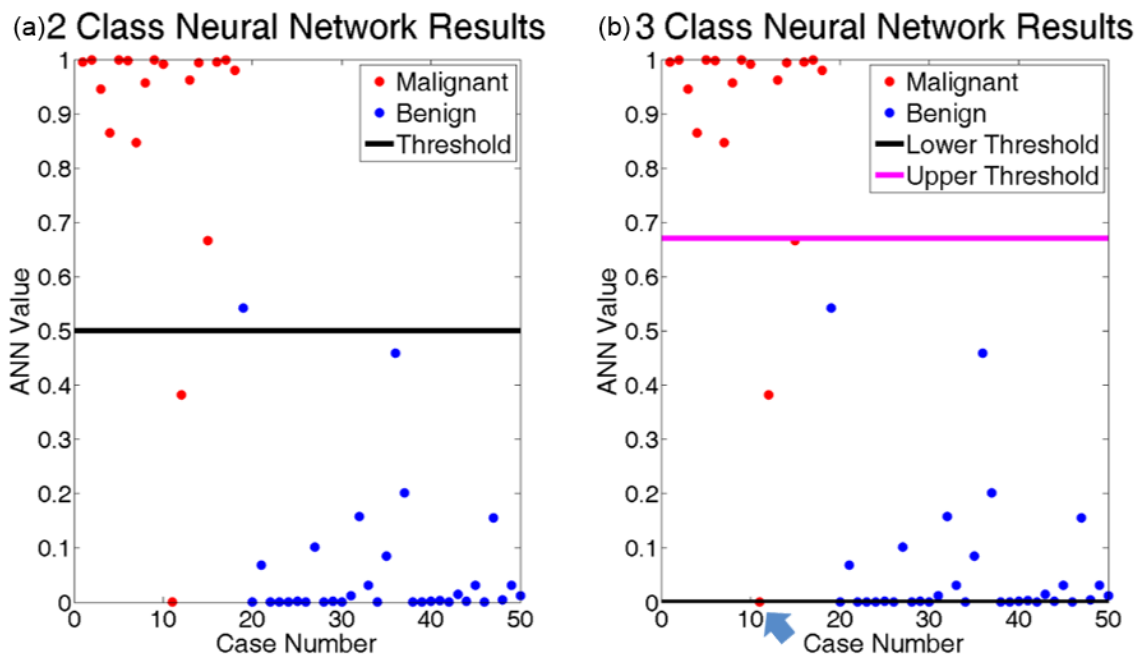

Figure 20. The neural network values for the N+50%+G experiment. The values of the network are plotted against the case number; correct diagnosis is indicated by the color of the point. The threshold(s) is(are) shown in black and magenta. (a) shows a single threshold/two class division. (b) shows the two threshold/three class division. Due to case 11 (see blue arrow in (b)), the lower threshold is set at 0.

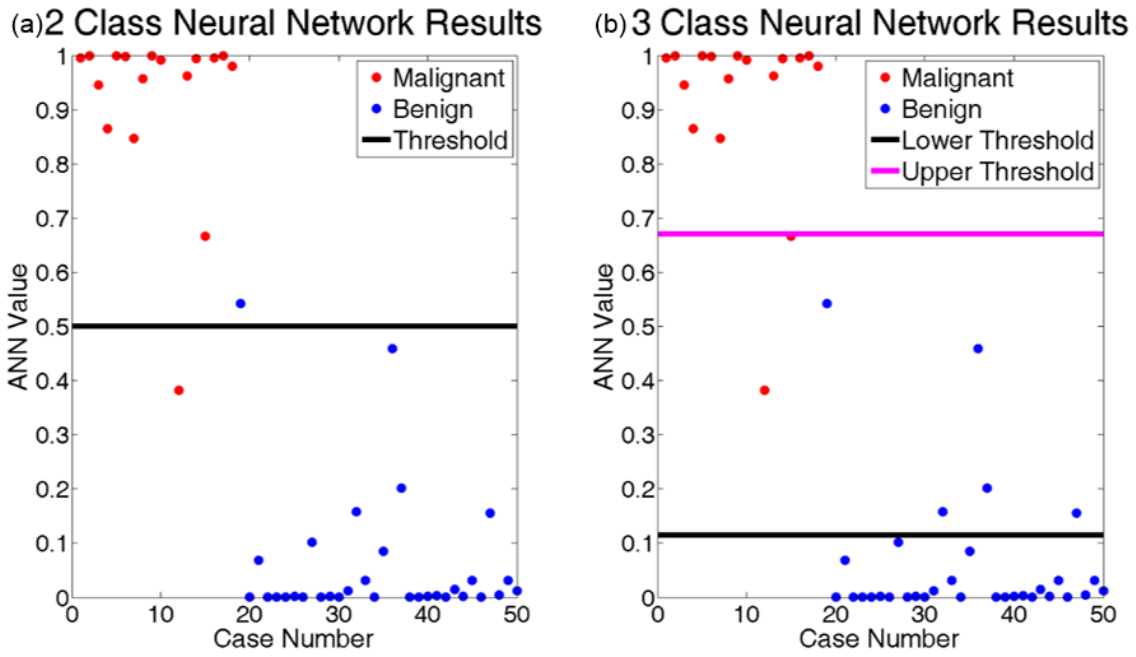

Figure 21. The neural network values for the N+50%+G experiment, excluding case 11. Upon its removal, the indeterminate class is much narrower, containing 7 (2 malignant, 5 benign) cases.

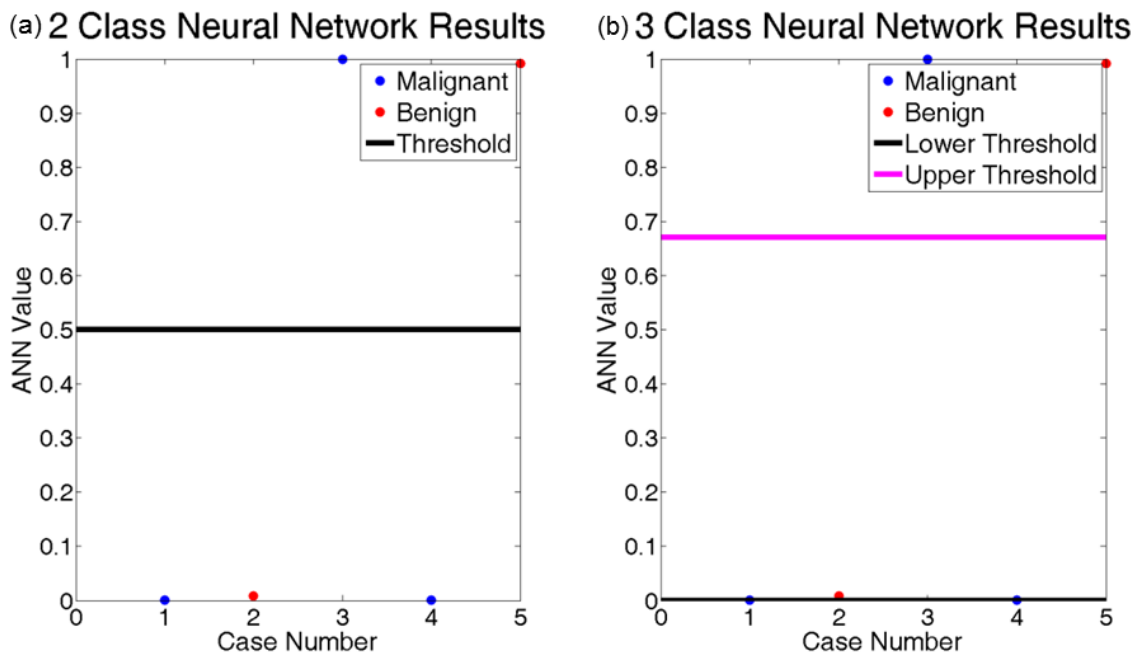

Figure 22. Research test cases for (a) binary thresholding and (b) three-class partitioning using the thresholds from all 50 cases. One benign nodule is misclassified in both (a) and (b), and one malignant nodule is misclassified in (a) and placed in the indeterminate category in (b).

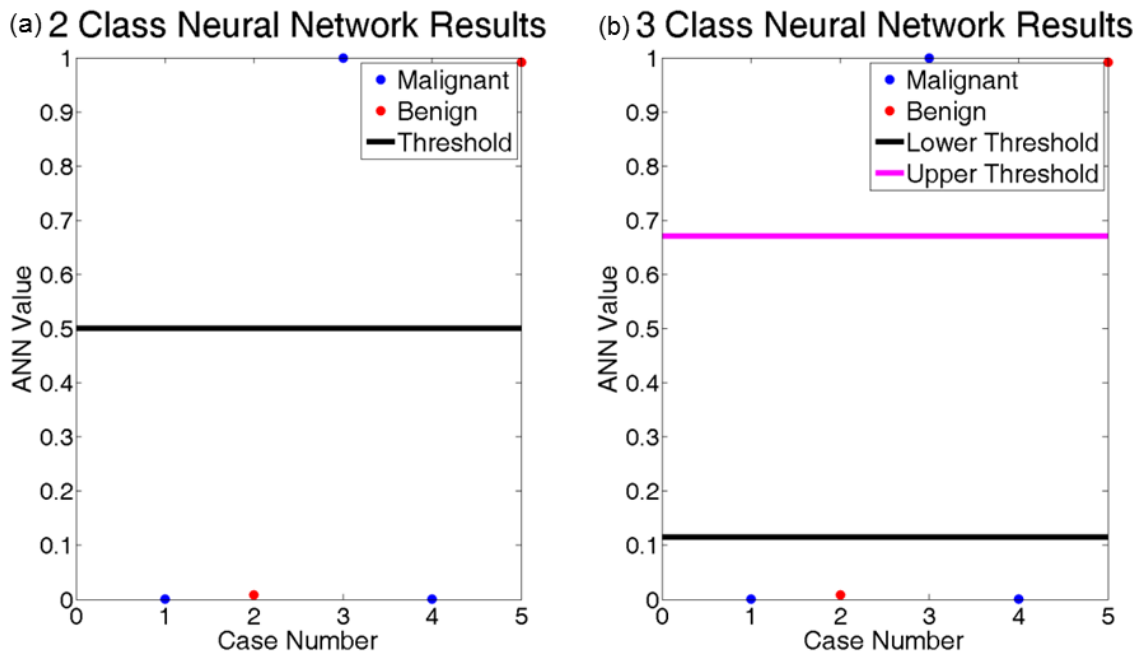

Figure 23. Research test cases for (a) binary thresholding and (b) three-class partitioning using the thresholds from the 49 cases. One malignant and one benign nodule are misclassified in both (a) and (b).

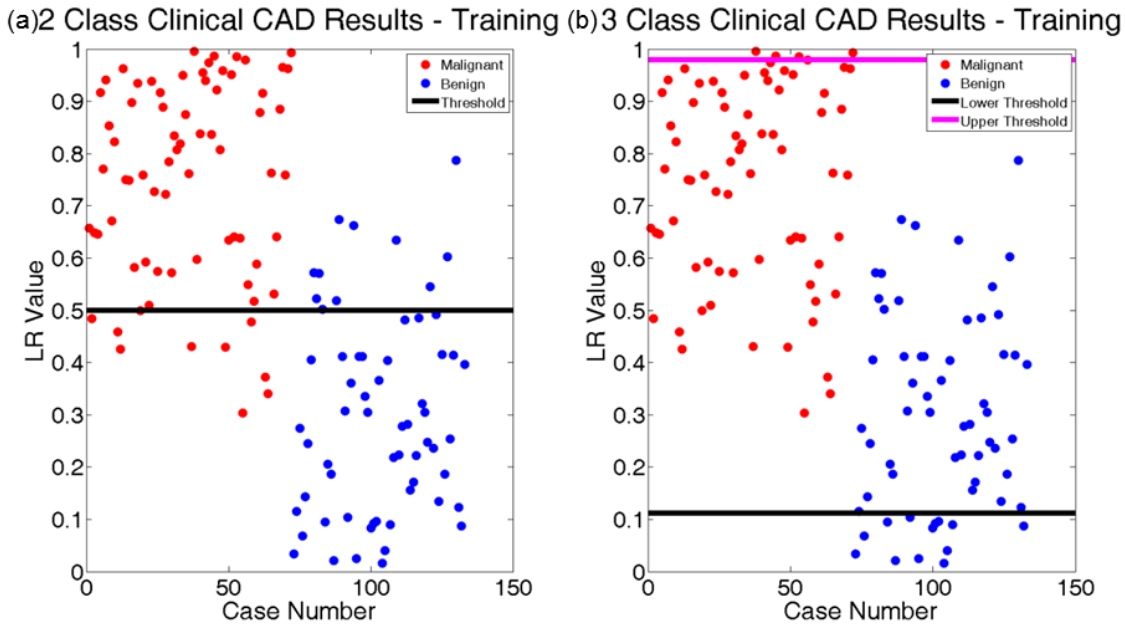

Figure 24. The logistic regression (LR) values for the N+50%+R experiment. The values of the LR are plotted against the training cases; correct diagnosis is indicated by the color of the point. The threshold(s) is(are) shown in black and magenta. (a) shows a single threshold/two class division. (b) shows the two threshold/three class division.

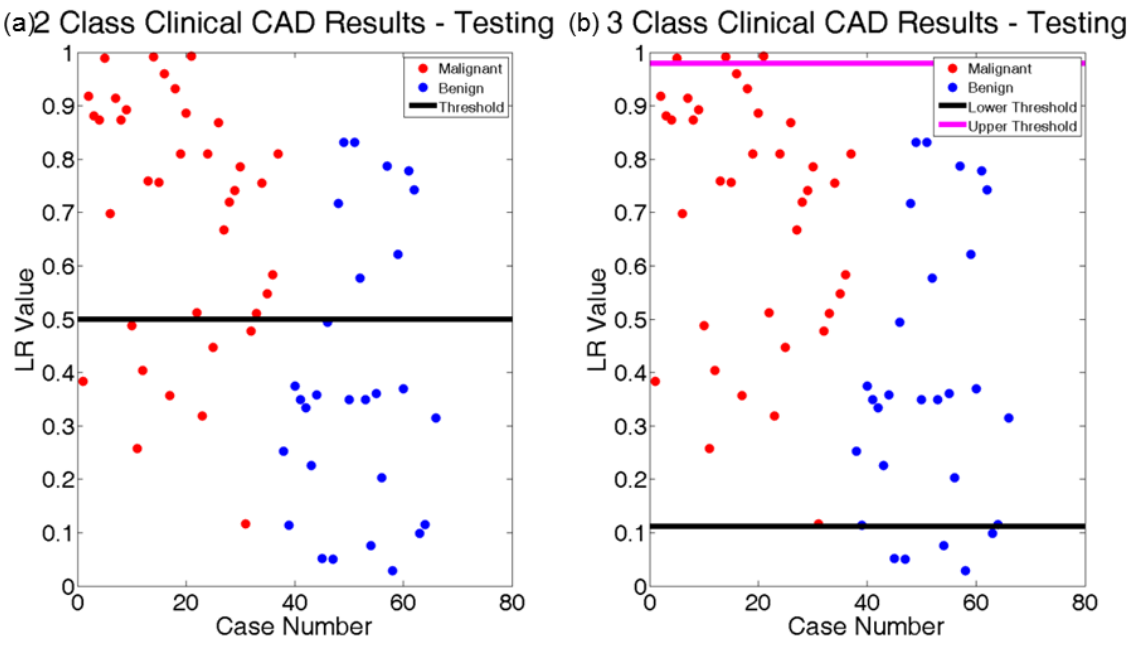

Figure 25. The logistic regression (LR) values for the N+50%+R experiment. The LR values are plotted against the testing cases, and the thresholds from the training set are applied in (b). No misclassifications occur.

CHAPTER VII

LONGITUDINAL CLINICAL IMPACT EXPLORATION

7.1 Rationale

According to the NCCN Guidelines, the detection of a pulmonary nodule in a low-dose Chest CT scan frequently results in repeated CT imaging at some interval, often 1-3 months, to assess nodule growth and change in appearance. These characteristics are subjective, relying on the reader's impressions to determine whether the nodule appears stable and continued CT monitoring is sufficient, or if the nodule requires further invasive evaluation to determine whether malignancy is present. In Aims 1 and 2, we identified QCT features that are useful for the classification of pulmonary nodules. For this study, we seek to explore the potential clinical impacts of the CAD tools incorporating parenchymal features by assessing the stability of nodule prediction.

7.2 Research Cohort

The longitudinal trends in nodule prediction for the research-driven CAD tool were investigated using a subset of the research cohort with more than one CT scan captured during NLST or COPDGene. In order to improve nodule follow-up, a logical trend in nodule prediction is desired, from a prediction of benign or indeterminate, progressing to more malignant as the disease develops and the nodule invades the surrounding parenchyma.

7.2.1 Methods

From the research cohort, 9 participants (3 malignant, 6 benign) with multiple CT scans present from the NIH trials were identified. NLST participants had up to three CT scans (one prevalence and two incidence scans), depending on when a nodule was identified. COPDGene participants without lung cancer had up to two CT scans (Phase I and Phase II 5 years after the initial timepoint). All timepoint CT scans were processed

through the optimized research CAD tool (N+50%+G); the nodules and the surrounding parenchyma were segmented, their 50% parenchyma masks generated, and the features used in classification (see Table 8) were extracted. These scans were then classified using a leave-one-case-out approach. From the original research cohort of 50 cases, the participant being classified in this longitudinal trial was left out of training; hence, 49 cases were used to train the ANN. After training, all CT scans of the left-out participant were processed through the trained classifier. This leaving-out procedure was repeated for all 9 subjects. Each training and testing iteration was repeated 10 times to obtain a better estimate of performance due to the random initialization of the ANN.

After acquiring ANN output estimates, the thresholds from *6.2.2 Results* for two class and three class classification were applied to the average value for each nodule. The trends in CAD diagnoses were plotted and visually assessed for illogical diagnosis transitions (malignant to benign).

7.2.2 Results

Of the 9 participants, 6 (3 malignant, 3 benign) were NLST participants, and 3 (benign) were from COPDGene. Each participant had two CT scans present; the NLST participants had one year between time points whereas the COPDGene participants had five years between timepoints. For the NLST nodules, the CT scan closest to the diagnosis date had been used in the development of the research CAD tool. For the COPDGene cases, the Phase I scan had been used in the CAD development process. When the two-class approach was used, one of the malignant nodules was initially diagnosed as malignant, and later diagnosed as benign (Figure 26).This case had a similar trend during the three class, two threshold approach; it was first diagnosed as malignant and later diagnosed as indeterminate for the 50-case threshold (lower threshold of 0, Figure 26) and benign for the 49-case threshold (lower threshold of 0.115, Figure 27).

The remainder of the cases show logical transitions from benign to indeterminate to malignant or remain stable in their predicted classification.

7.3 Clinical Cohort

7.3.1 Methods

From the clinical cohort, a subcohort of patients diagnosed with either a malignant or benign pulmonary nodule with multiple retrospective CT scans was randomly identified. Their CT scans were collected and processed through the optimized clinical CAD tool (N+50%+R, see list of features in Table 9).

After optimized feature extraction had occurred, the nodules were processed through the N+50%+R LR model trained with the training set of 133. The LR predictions were processed through the three class, two threshold approach and the trends in nodule prediction were examined and compared with the radiologist impressions.

7.3.2 Results

From the clinical cohort, 24 patients with a solitary pulmonary nodule (12 malignant, 12 benign) with multiple CT scans were identified. The malignant nodules were imaged 3.3±1.4 times on average over 1.3±1.8 years before diagnosis, and the benign nodules were followed for an average of 2.7±1.0 CT scans over 1.0±1.5 years.

Figure 28 illustrates the 24 clinical cases with their actual diagnoses. Each CT scan has been demarcated by the radiologist's impression of the nodule. In several cases, especially 9, 11, 13, and 23, variability in the radiologist's impression is present across the nodule's imaging timeline. When processed through the optimal clinical CAD tool (N+50%+R LR), these cases are either classified as indeterminate or show a trend from benign to indeterminate (Figure 29). While no nodules were classified as malignant, four nodules were classified as definite benign. This reduces the number of benign cases referred to follow-up from 12 to 8, a reduction of 33%.

As some of the subcohort were used for training of the LR model, those cases that were in the testing set were isolated and examined. These cases included 3 benign (cases 3, 5, and 7) and 5 malignant (cases 15, 17, 18, 19, and 22). Similarly to when all 24 were assessed, 1 of the 3 benign nodules was classified as definite benign from the initial timepoint, yielding a reduction of 33% of benign cases to follow-up.

7.4 Discussion

In this subproject, we sought to explore the clinical impact a three-class approach could have on the follow-up procedure for CT-identified pulmonary nodules. By identifying patients with multiple CT scans in our two cohorts, we collected two longitudinal subcohorts with which to visualize the trends in CAD predictions. In both groups, the use of three classes would have eliminated repeated scanning for benign nodules. As most nodules identified with CT are benign, this has the potential to improve healthcare costs and reduce patient risks associated with invasive procedures and repeated imaging.

In the research longitudinal subcohort, the tighter CT protocol allows for more confidence that differences seen within features and CAD predictions are due to physiological changes in the nodule and its surrounding parenchyma. 5 of the 6 benign nodules were correctly classified as definite benign within the three-class thresholding scheme constructed from the 49 research cases.

In the 24-case clinical cohort, 4 of the 12 benign nodules were labeled as definite benign at first CT scan. The potential gains from this are undersold due to the use of a balanced (12 benign and 12 malignant) division of diagnoses. The NLST found 97% of lesions identified by CT were benign with additional follow-up. The reduction of clinical follow-up on 33% of these lesions would have a huge clinical benefit: reduction of repeated imaging, invasive biopsies and procedures, reduced patient stress, and reduced healthcare costs.

In both CAD tools, nodule size was not incorporated in the decision of nodule prediction. However, clinically, nodule growth is a major indication of malignancy. A limitation of our developed approach in the clinical cohort was the lack of transition from indeterminate to malignant prediction for the malignant cases. Obviously, limiting follow-up imaging procedures and obtaining a swift diagnosis for malignant cases is also very important. Modifications to the algorithm to include longitudinal change in features should be explored in the future to add value for this cohort. However, for the research subcohort, the prediction trends seen when a tighter protocol is used suggest nodule size may not be as important as other changes quantified. In this preliminary work, each CT ROI was assessed as an independent case through the clinical CAD tool. The positive performance on this challenging dataset shows the promise of longitudinal nodule classification and is worth further development by incorporating longitudinal change in features.

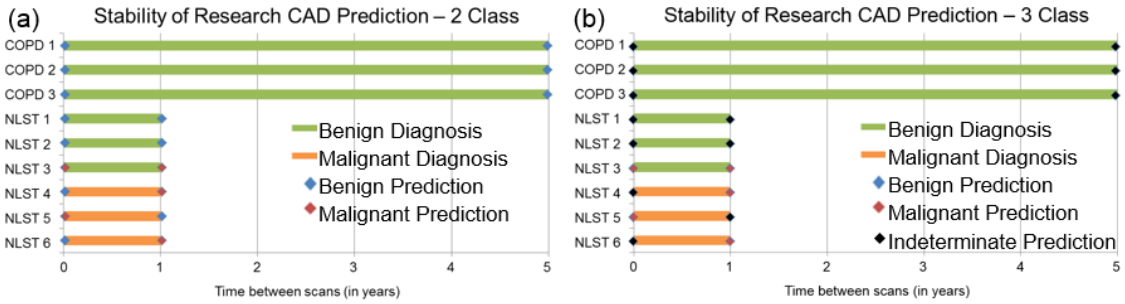

Figure 26. Two and three class (built with 50 cases) CAD predictions for the research stability subcohort. 3 malignant (orange lines) and 6 benign (green lines) cases are shown. Data points represent individual CT scans, with the length of the line indicating the time between scans. The color of the data point indicates the predicted CAD diagnosis: malignant and benign for the two class approach (a) and malignant, benign, or indeterminate for the three class approach (b).

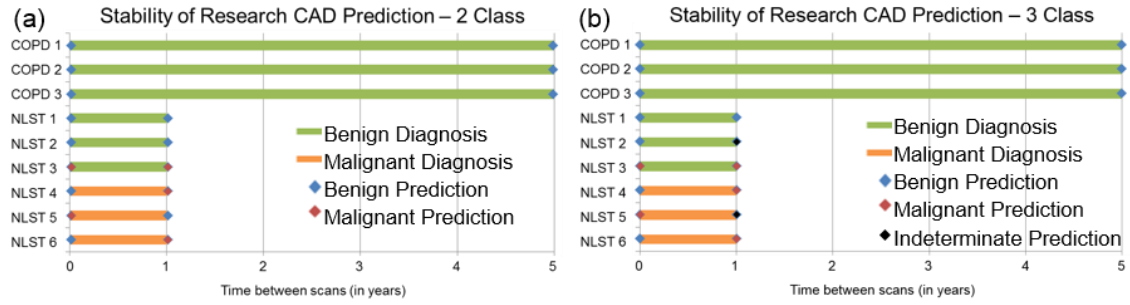

Figure 27. Two and three class (built with 49 cases) CAD predictions for the research stability subcohort. Four of the six benign nodules are classified as definite benign when the three class approach built with 49 cases (b) is used, compared to the three class approach when all 50 cases were used to set the thresholds (Figure 26).

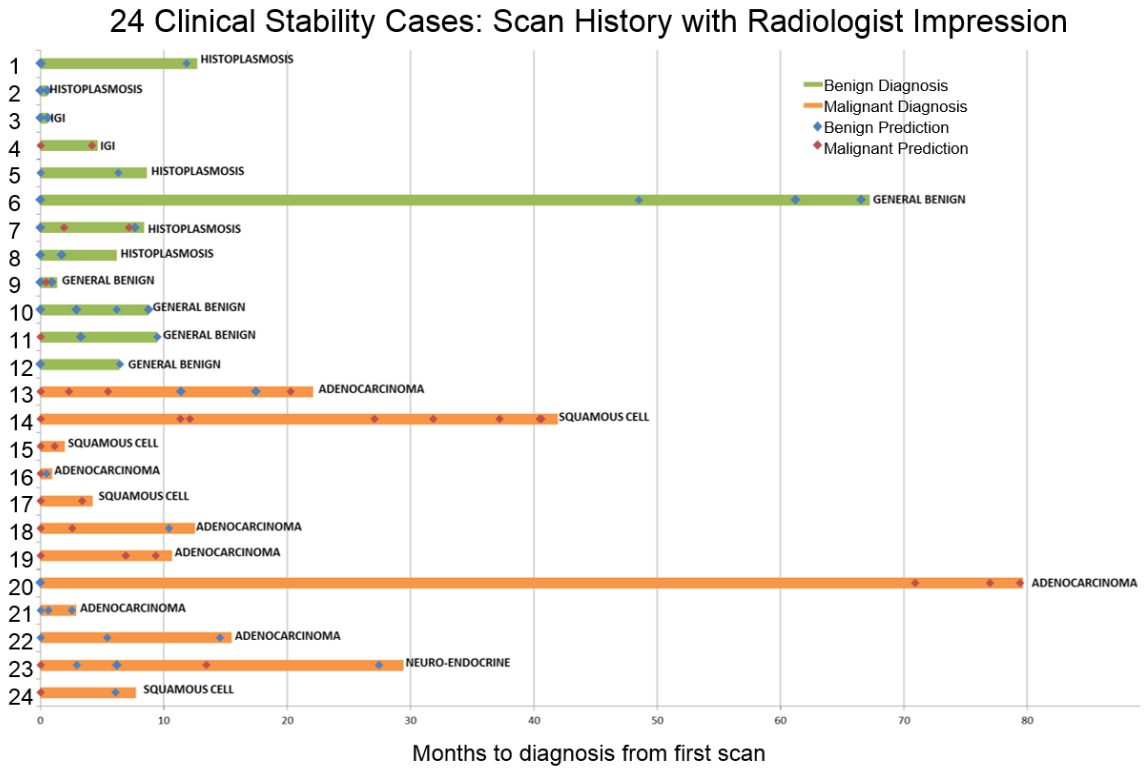

Figure 28. Scan history for clinical subcohort. 12 malignant (orange lines) and 12 benign (green lines) cases are shown, with histology diagnosis listed. The length of the line represents the time from identification in CT to time of diagnosis. Data points represent individual CT scans, with the radiologist's impression noted by color (red: malignant impression, blue: benign impression).

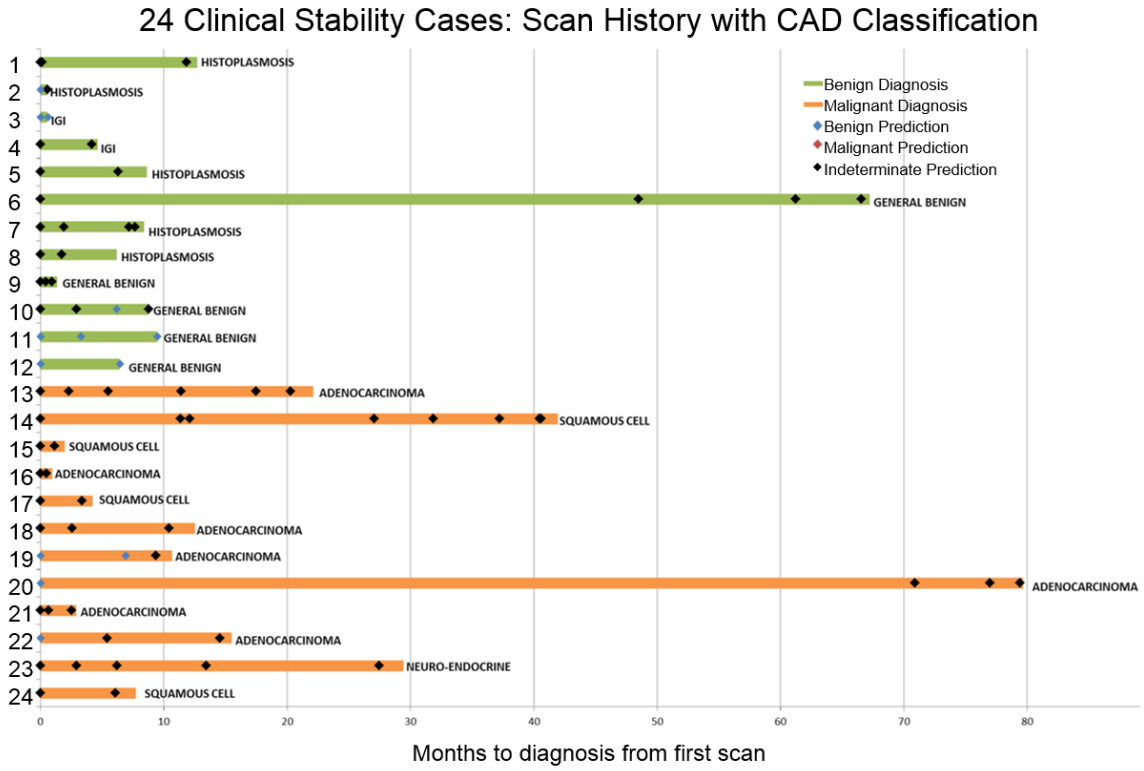

Figure 29. Scan history with CAD classification for the clinical subcohort. The CT scans for 24 clinical subjects were processed through the optimized N+50%+R LR classifier. 4 of the 12 benign nodules were classified as definite benign, reducing follow-up for these participants.

CHAPTER VIII

CONCLUSION

As low-dose CT screening for lung cancer becomes wide-spread, the management of CT-identified lung nodules will become an even greater clinical dilemma. CAD tools have the potential to improve CT specificity by acting as a second reader to radiologists, however, traditional models rely solely on nodule characters. We hypothesize that improvement will be seen in the quantitative CT (QCT) CAD classification of lung nodules by including the lung parenchyma surrounding the nodule in the region of interest. We explored our hypothesis through three specific aims:

Aim 1: *Develop and test a CAD tool that utilizes QCT features derived from the surrounding parenchyma in both a high CT resolution research cohort and a retrospective clinical cohort with greater variance in CT resolution and acquisition parameters*. We developed two CAD tools: a research-driven, lung quantification tool and a clinically-focused, lung cancer screening tool robust to variations in protocol. Both CAD tools utilized parenchymal features and performed better than CAD tools trained solely with nodule features.

Aim 2: *Determine the impact of including parenchymal features by identifying the optimal amount of parenchyma surrounding a nodule to include in the CAD tool.* We further optimized the developed CAD tools be identifying the optimal amount of parenchyma to assess during feature extraction. Both CAD tools benefited from including parenchyma within a distance of 50% of the nodule's diameter.

Aim 3: *Explore the potential clinical impact of a CAD tool that includes parenchymal features in providing an early prediction of malignancy and minimizing follow-up imaging and procedures in the cohort with benign nodules.* Two exploratory studies were performed. A procedure to develop a three class (definite malignant, definite benign, and indeterminate) approach was found to reduce the number of follow-up

procedures necessary through NCCN Guidelines in both the research and clinical cohorts. Additionally, the longitudinal analysis on a subset of nodules in each cohort showed potential in reducing follow-up procedures.

CHAPTER IX

FUTURE WORK

Several projects are planned in expanding this CAD tool. They vary in range from expanding datasets, extending feature extraction, improving classification, and incorporating an automated segmentation method.

9.1 Application to Lung Cancer Screening Protocol

At the University of Iowa Hospitals and Clinics, a preventative screening protocol is being designed to image those patients qualifying for lung cancer screening under USPSTF guidelines. These patients will consent to having their CT images and any resulting pathology results used for research purposes. This provides the opportunity for a CAD tool to be designed for a tighter controlled population and low-dose protocol by using the CAD modules described in this work.

9.2 Expanded Variability of Cases

The current dataset contains primarily solid nodules. In the future, an increased percentage of sub-solid nodules will be added to the training and testing datasets. Ground glass nodules (GGN) are defined as localized regions of increased attenuation through which it is possible to visualize normal pulmonary structures (parenchyma, airways and vessels). The category of sub-solid nodules encompasses purely GGN and partially solid GGN. This class of nodules present challenges in the areas of quantification due to some ambiguity in defining the nodule boundary. It is likely that the expanded parenchymal feature sets described in this thesis will be highly suited to the analysis of sub-solid nodules.

9.3 Feature Extraction Methods

9.3.1 Rubber Band Straightening Transform

One limitation of the current set of feature extraction scripts lies in the RBST function. At present, the function is suitable for nodules whose centroid lies within their boundaries. For nodules with great concavities or nonsolid centers, this function is unable to properly assess the nodule's border, as the algorithm identifies one crossing of the border. Expanding this feature set would be necessary to interpret a greater variety of nodule types.

9.3.2 Regional Lung Analysis

In the research cohort, features extracted from the lobe containing the nodule were found to aid in classification. However, the clinical scans cannot be processed through Apollo due to the scan parameter limitations, including slice thickness. A module to look at lobar measures in the clinical cohort could improve pulmonary nodule prediction further.

9.4 Feature Selection and Classification

While lasso-penalized logistic regression was found to be the best feature selection and classification method for the clinical cohort, another feature selection method shows promise given the size of the clinical cohort. AdaBoost methods iteratively combine "weak" classifiers, constructed from individual features, to generate a single, strong classifier [48]. This is performed by iteratively increasing the influence misclassified cases have on the next round of classifier selection. As each classifier is constructed using a single feature, this serves as a feature selection tool. The training set would be used to train the weak classifiers and determine the optimal number of weak classifiers to use in the final classification schema, while the testing set would be available for unbiased testing.

9.5 Refinement of the Three-Class Approach

The promising results from *CHAPTER VI IMPLEMENTING A THREE-CLASS APPROACH* showed the potential of the CAD tools to reduce the number of nodules sent for follow-up, as well as a reduction in time for patients with malignant nodules to undergo treatment. The current method relies on classifier outputs to determine the thresholds for definite malignant, definite benign, and indeterminate classes. An optimal approach would incorporate a cost function into the classification training phase to classify nodules directly into one of the three classes.

9.6 Longitudinal Explorations

The analysis of trends for the longitudinal dataset suggest these CAD tools could have considerable impact on the management of CT-identified lung nodules. By analyzing each participant's CT scan as an independent ROI, the relative change in the nodule was left unaccounted. We predict that the incorporation of relative change features into the CAD tools would improve the performance of the CAD tools when nodules are re-imaged and the next decision in nodule management (repeat imaging, biopsy, or surgery) is needed. The nodule's growth rate, change in shape (represented by sphericity and border assessment), and change in parenchyma texture (as the nodule infiltrates and influences the surrounding parenchyma) would be included as features. These features, as well as the features outlined in *4.1.1.1 Feature Extraction* would be processed through feature selection and classification in order to develop a more robust longitudinal CAD tool.

APPENDIX

Table A 1. Categorical information for binning the reconstruction kernels into six types to check for confounders within the cohort; none were identified.

Category	**Kernels Included:**
Standard	B, B30f, B30s, B31f,B31s,B40s,B41f,B41s, Standard
Soft	B20f, B20s
Sharp	B46f, B70f, B70s
Standard without beam hardening correction	FC10,FC12,FC13,FC14,FC17,FC18
Iterative	I31f, I41f
Toshiba standard	FC01, FC02, FC03

Table A 2. List of participants from National Lung Screening Trial (NLST) and COPDGene used in the research cohort. The imaging timepoint used in this study is indicated, with year 0, 1, or 2 for NLST participants and phase 1 or 2 for COPDGene participants.

Subject Identification Number	Malignant Nodules Timepoint	Subject Identification Number	Benign Nodules Timepoint
National Lung Screening Trial		***National Lung Screening Trial***	
14983	0	2674	2
16162	0	3566	1
5118	2	3630	2
7268	2	4061	2
12077	1	4697	1
10072	1	15104	1
		16092	1
		6755	1
COPDGene		***COPDGene***	
21700T	1	13856C	1
21448J	1	18786G	1
25721V	1	12035J	1
21463F	1	10157N	1
10129I	1	14652R	1
18904M	1	10283S	1
21450W	1	12272X	1
13250S	1	14352F	1
16774P	1	20489S	1
17450U	1	10213X	1
15836G	1	10287A	1
13092A	1	10695P	1
10309K	1	10750X	1
24060Y	1	12556L	1
		13323T	1
		10127E	1
		10214Z	1
		11004T	1
		11824F	1
		12056R	1
		14290J	1
		20226M	1
		20545C	1
		19756C	1
		10025W	1
		24757L	1
		11720T	1

Table A 3. List of equations for less common features used in Feature Extraction. A complete description can be found in [1].

Feature	Equation
Entropy, H(X)	$H(X) = -\sum_{k=1}^{n} p(x_k)\log_2(p(x_k))$ $p(x_k)$: probability of an outcome
Kurtosis	$Kurtosis = \frac{\sum_{i=1}^{N}(Y_i-\bar{Y})^4}{(N-1)s^4}$ Y: individual intensity value, HU $\bar{Y}$: mean intensity value, HU N: umber of voxels in the region s: standard deviation of intensity
Skewness	$Skewness = \frac{\sum_{i=1}^{N}(Y_i-\bar{Y})^3}{(N-1)s^3}$ Y: individual intensity value, HU $\bar{Y}$: mean intensity value, HU N: number of voxels in the region s: standard deviation of intensity
Sphericity	$Sphericity = \frac{\pi^{\frac{1}{3}}(6V)^{\frac{2}{3}}}{SA}$ V: volume of the nodule SA: surface area of the nodule
Effective Radius, r	$V = \frac{4}{3}\pi r^3 \rightarrow r = \left(\frac{3V}{4\pi}\right)^{\frac{1}{3}}$ r: radius of a sphere V: volume of a sphere

Table A 4. Complete list of features extracted during research CAD development.

Features Extracted for Research CAD Tool
Intensity Features
Nodule Mean Intensity
Nodule Variance
Nodule Maximum Intensity
Nodule Minimum Intensity
Nodule Median Intensity
Nodule Full Width Half Maximum
Nodule Entropy
Nodule Kurtosis
Nodule Skewness
Parenchyma Mean Intensity
Parenchyma Variance
Parenchyma Maximum Intensity
Parenchyma Minimum Intensity
Parenchyma Median Intensity
Parenchyma Full Width Half Maximum
Parenchyma Entropy
Parenchyma Kurtosis
Parenchyma Skewness
Shape Features
Sphericity
Absolute Difference in Mean in Boundary
Absolute Difference in Variance in Boundary
Absolute Difference in Kurtosis in Boundary
Absolute Difference in Skewness in Boundary
Absolute Difference in Range in Boundary
Effective Radius
RECIST Maximum Diameter
Border Features
Mean of the Border's Width in Physical Space
Mean of the Border's Width in Pixel Space
Standard Deviation of the Border's Width in Physical Space
Standard Deviation of the Border's Width in Pixel Space
Mean Intensity Change Across the Border
Variance in Intensity Change Across the Border
Texture Features
34 3D Nodule Texture Mean Features
34 3D Nodule Texture Variance Features
34 3D Nodule Texture Kurtosis Features
34 3D Nodule Texture Skewness Features
34 3D Parenchyma Texture Mean Features
34 3D Parenchyma Texture Variance Features
34 3D Parenchyma Texture Kurtosis Features
34 3D Parenchyma Texture Skewness Features

Table A 5. Complete list of features extracted during clinical CAD development.

Features Extracted for Clinical CAD Tool
Risk Factors
Age
Race
Sex
Pack Years
Lobe
Intensity Features
Nodule Mean Intensity
Nodule Variance
Nodule Maximum Intensity
Nodule Minimum Intensity
Nodule Median Intensity
Nodule Full Width Half Maximum
Nodule Entropy
Nodule Kurtosis
Nodule Skewness
Parenchyma Mean Intensity
Parenchyma Variance
Parenchyma Maximum Intensity
Parenchyma Minimum Intensity
Parenchyma Median Intensity
Parenchyma Full Width Half Maximum
Parenchyma Entropy
Parenchyma Kurtosis
Parenchyma Skewness
Shape Features
Sphericity
Absolute Difference in Mean in Boundary
Absolute Difference in Variance in Boundary
Absolute Difference in Kurtosis in Boundary
Absolute Difference in Skewness in Boundary
Absolute Difference in Range in Boundary
Effective Radius
RECIST Maximum Diameter
Border Features
Mean of the Border's Width in Physical Space
Mean of the Border's Width in Pixel Space
Standard Deviation of the Border's Width in Physical Space
Standard Deviation of the Border's Width in Pixel Space
Mean Intensity Change Across the Border
Variance in Intensity Change Across the Border
Texture Features
14 2D Nodule Texture Mean Features
14 2D Nodule Texture Variance Features
14 2D Nodule Texture Kurtosis Features
14 2D Nodule Texture Skewness Features
14 2D Parenchyma Texture Mean Features
14 2D Parenchyma Texture Variance Features
14 2D Parenchyma Texture Kurtosis Features
14 2D Parenchyma Texture Skewness Features

Table A 6. Global features extracted during N+50%+G CAD development with the research cohort.

Global Features Extracted for Research CAD Tool
Whole Lung Features
Mean Intensity
Standard Deviation of Intensity
Kurtosis
Skewness
Percent Emphysema (Percent Below -950 HU in inspiration)
Percent Below -910 HU
Percent Gas Trapping (Percent Below -856 HU in expiration)
Percent Above 0 HU
Hounsfield Value at 15th Percentile
Total Volume
Tissue Volume
Air Volume
Lung Specific Features
Mean Intensity
Standard Deviation of Intensity
Kurtosis
Skewness
Percent Emphysema (Percent Below -950 HU in inspiration)
Percent Below -910 HU
Percent Gas Trapping (Percent Below -856 HU in expiration)
Percent Above 0 HU
Hounsfield Value at 15th Percentile
Total Volume
Tissue Volume
Air Volume
Lobe Specific Features
Mean Intensity
Standard Deviation of Intensity
Kurtosis
Skewness
Percent Emphysema (Percent Below -950 HU in inspiration)
Percent Below -910 HU
Percent Gas Trapping (Percent Below -856 HU in expiration)
Percent Above 0 HU
Hounsfield Value at 15th Percentile
Total Volume
Tissue Volume
Air Volume
Third Specific Features
Mean Intensity
Standard Deviation of Intensity
Kurtosis
Skewness
Percent Emphysema (Percent Below -950 HU in inspiration)
Percent Below -910 HU
Percent Gas Trapping (Percent Below -856 HU in expiration)
Percent Above 0 HU
Hounsfield Value at 15th Percentile
Total Volume
Tissue Volume
Air Volume

Table A 7. Summary of features selected and performances for the training and testing sets of the clinical cohort for the QCT and risk factors (N+P+R).

Selected Features with LASSO	Training AUC	Testing AUC
QCT Features + Risk Factors (N+P+R)		
QCT - 2D textures	0.93	0.81
Lobe		
Age		
Sex		
Pack Years		
Nodule Full Width Half Maximum		
Nodule Entropy		
Parenchyma Mean Intensity		
Absolute Difference in Range in Boundary		
2D Nodule Texture Variance 10		
2D Nodule Texture Skewness 3		
QCT - 3D textures	0.92	0.81
Lobe		
Age		
Sex		
Pack Years		
Nodule Full Width Half Maximum		
Nodule Entropy		
Parenchyma Mean Intensity		
Absolute Difference in Range in Boundary		
Effective Radius		
3D Nodule Texture Variance 18		
3D Parenchyma Texture Variance 1		
QCT - 3D textures, no size features	0.92	0.80
Lobe		
Age		
Sex		
Pack Years		
Nodule Maximum Intensity		
Nodule Entropy		
Nodule Full Width Half Maximum		
Parenchyma Mean Intensity		
Absolute Difference in Range in Boundary		
3D Nodule Texture Variance 18		
3D Parenchyma Texture Variance 1		
3D Parenchyma Texture Kurtosis 34		
QCT - 2D and 3D textures	0.92	0.81
Lobe		
Age		
Sex		
Pack Years		
Nodule Full Width Half Maximum		
Nodule Entropy		
Parenchyma Mean Intensity		
Absolute Difference in Mean in Boundary		
Absolute Difference in Range in Boundary		
3D Parenchyma Texture Variance 1		
2D Nodule Texture Variance 10		
2D Nodule Texture Skewness 3		

Table A 8. Summary of features selected and performances for the training and testing sets of the clinical cohort for the QCT features (N+P).

Selected Features with LASSO	Training AUC	Testing AUC
QCT Features (N+P)		
QCT - 2D textures	0.79	0.68
Nodule Entropy		
Parenchyma Mean Intensity		
Parenchyma Minimum Intensity		
Absolute Difference in Mean in Boundary		
Absolute Difference in Variance in Boundary		
Effective Radius		
2D Nodule Texture Skewness 3		
QCT - 2D textures, no size features	0.76	0.59
Nodule Entropy		
Parenchyma Mean Intensity		
Parenchyma Minimum Intensity		
Absolute Difference in Mean in Boundary		
Mean of the Border's Width in Pixel Space		
QCT - 3D textures	0.81	0.69
Nodule Full Width Half Maximum		
Nodule Entropy		
Parenchyma Mean Intensity		
Parenchyma Minimum Intensity		
Parenchyma Full Width Half Maximum		
Absolute Difference in Mean in Boundary		
Absolute Difference in Variance in Boundary		
Absolute Difference in Range in Boundary		
Effective Radius		
3D Parenchyma Texture Variance 1		
QCT - 2D and 3D textures	0.80	0.68
Nodule Entropy		
Parenchyma Mean Intensity		
Parenchyma Minimum Intensity		
Absolute Difference in Mean in Boundary		
Absolute Difference in Variance in Boundary		
Effective Radius		
3D Parenchyma Texture Variance 1		
2D Nodule Texture Skewness 3		
QCT - 2D and 3D textures, no size and 3D textures, no size	0.77	0.59
Nodule Entropy		
Parenchyma Mean Intensity		
Parenchyma Minimum Intensity		
Absolute Difference in Mean in Boundary		
Mean of the Border's Width in Pixel Space		
3D Parenchyma Texture Variance 1		

Table A 9. Summary of features selected and performances for the training and testing sets of the clinical cohort for the N+50%+R CAD tool when contrast-enhanced scans alone are used.

Amount of Parenchyma	Clinical Contrast-enhanced N+P+R LR Classifier		
	QCT Features + Risk Factors (N+P+R)	Training AUC	Testing AUC
50%	**Intensity** Nodule variance Nodule full-width at half maximum Nodule entropy Parenchyma mean intensity Parenchyma median intensity **Shape** Absolute difference range in boundary **Texture** Nodule variance 2D texture Nodule skewness 2D texture Parenchyma variance 2D texture **Risk Factors** Lobe location Age Sex Pack-years	0.95	0.79

Table A 10. Means and standard deviations (St. Dev.) for features selected by stepwise feature selection for the research CAD classifiers (N, N+P, N+50%, N+(50-10)%, and N+50%+G).

Regions of Interest	Malignant		Benign		P-values
	Mean	St. Dev.	Mean	St. Dev.	
Nodule Features: N					
Nodule Variance	35956.0	24240.9	77685.0	130626.0	1.1E-4
Nodule Kurtosis TEM17	3279.2	7143.9	614.9	963.0	1.3E-2
Mean of Border	9.0	5.1	6.5	2.8	7.9E-3
Nodule Kurtosis TEM2	3174.5	7209.0	486.6	536.0	1.0E-2
Nodule Kurtosis TEM16	3235.3	7170.1	556.5	724.5	1.7E-2
Nodule & Parenchyma Features: N+P					
Parenchyma Variance	16163.3	8759.2	8360.7	5496.1	1.1E-4
Nodule Minimum HU	-790.8	94.5	-841.4	67.6	5.4E-3
Nodule Kurtosis TEM1	3070.0	7080.3	461.0	506.1	7.0E-3
Nodule Kurtosis TEM15	3154.4	6993.4	511.7	580.8	1.5E-2
Nodule & Parenchyma Features: N+50%					
Parenchyma Entropy	8.5	0.4	8.0	0.4	2.8E-5
Nodule Median HU	-134.1	154.6	-230.7	207.1	2.0E-2
Abs. Difference Range	0.7	0.2	0.6	0.2	2.0E-2
Parenchyma Variance	20647.7	11828.8	12498.7	9280.6	4.0E-3
Nodule Variance	35956.0	24240.9	77685.0	130626.0	1.1E-4
Nodule & Parenchyma Features: N+(50-10)%					
Parenchyma Skewness TEM29	-3.2	14.4	-0.5	5.1	7.0E-4
Abs. Difference Range	0.7	0.2	0.6	0.2	2.0E-2
Parenchyma Skewness TEM34	-1.5	14.5	-0.2	5.0	1.3E-2
Nodule Variance	35956.0	24240.9	77685.0	130626.0	1.1E-4
Parenchyma Entropy	8.3	0.4	7.8	0.4	3.8E-5
Nodule & Parenchyma Features: N+50%+G					
Parenchyma Entropy	8.5	0.4	8.0	0.4	2.8E-5
Lobe-specific Skewness	2.8	0.5	3.3	0.6	1.7E-3
Nodule Minimum HU	-790.8	94.5	-841.4	67.6	5.4E-3
Nodule Mean HU	-177.8	128.6	-242.1	209.5	1.3E-2
Lobe-specific Standard Deviation	152.4	12.6	144.2	13.2	3.7E-2

Table A 11. Means and standard deviations (St. Dev.) for the quantitative features selected by lasso-penalized logistic regression for the final clinical CAD classifier (N+50%+R). Two categorical variables, sex and nodule location, were also selected.

Regions of Interest	Malignant		Benign		P-values
	Mean	St. Dev.	Mean	St. Dev.	
Nodule & Parenchyma Features: N+50%+R					
Intensity					
Nodule Variance	33834.0	13987.6	36693.9	25322.4	0.72
Nodule Full-Width at Half Maximum	4.1E-2	3.3E-2	3.2E-2	2.0E-2	2.2E-2
Nodule Entropy	8.4	0.5	8.3	0.7	0.19
Parenchyma Median Intensity	-801.7	80.6	-763.0	105.7	7.5E-4
Shape					
Abs. Difference Range	0.6	0.3	0.7	0.3	0.41
Texture					
Nodule Variance 2D TEM10	2.2E-3	5.1E-3	2.8E-3	4.8E-3	0.97
Nodule Skewness 2D TEM3	0.7	13.5	-2.0	9.4	0.12
Parenchyma Variance 2D TEM4	1.5E-5	9.6E-5	1.1E-5	6.9E-5	0.22
Risk Factors					
Age	66	11	53	13	<1.0E-5
Pack-Years	38	32	11	17	<1.0E-5

REFERENCES

1. Dilger, S.K.N., *The use of surrounding lung parenchyma for the automated classification of pulmonary nodules*, in *Biomedical Engineering* 2013, University of Iowa.
2. Dilger, S.K.N., et al., *Improved pulmonary nodule classification utilizing quantitative lung parenchyma features.* Journal of Medical Imaging, 2015. **2**(4): p. 041004.
3. Howlader, N., et al., (eds). *SEER Cancer Statistics Review, 1975-2012*. Updated 18 Nov 2015, Accessed 15 Feb 2016; Available from: http://seer.cancer.gov/csr/1975_2012/sections.html.
4. Aberle, D.R., et al., *The National Lung Screening Trial: overview and study design.* Radiology, 2011. **258**(1): p. 243-53.
5. The National Lung Screening Trial Research Team, *Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening.* New England Journal of Medicine, 2011. **365**(5): p. 395-409.
6. *Lung Cancer (Small Cell)*, 2013, American Cancer Society.
7. Ganti, A.K. and D.E. Gerber, *Lung cancer*. 2013.
8. Bach, P.B., et al., *Screening for Lung Cancer: A review of the current literature.* Chest, 2003. **123**(1): p. 72-82.
9. The Prostate, Lung, Colorectal, and Ovarian (PLCO) Randomized Trial, et al., *Screening by chest radiograph and lung cancer mortality: the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial.* JAMA, 2011. **306**(17): p. 1865-73.
10. Moyer, V.A. *Screening for Lung Cancer: U.S. Preventive Services Task Force Recommendation Statement*. Annals of Internal Medicine, 2013. 1-11.
11. The Centers for Medicare & Medicaid Services, *Decision Memo for Screening for Lung Cancer with Low Dose Computed Tomography (LDCT) (CAG-00439N)*, 2015.
12. National Comprehensive Cancer Network, (eds). *NCCN Guidelines Version 1.2016 Lung Cancer Screening*. Updated Accessed 28th January 2016; Available from: http://www.nccn.org/professionals/physician_gls/pdf/lung_screening.pdf.
13. Revel, M.-P., et al., *Are Two-dimensional CT Measurements of Small Noncalcified Pulmonary Nodules Reliable?* Radiology, 2004. **231**(2): p. 453-458.
14. McNitt-Gray, M.F., et al., *Determining the variability of lesion size measurements from ct patient data sets acquired under "no change" conditions.* Translational Oncology, 2015. **8**(1): p. 55-64.
15. Burfeind Jr., W., et al., *Early Detection of Lung Cancer by a FISH-Based Sputum Test*, in *16th World Conference on Lung Cancer*2015: Denver, CO.
16. Molina, R., et al., *Assessment of a Combined Panel of Six Serum Tumor Markers for Lung Cancer.* American Journal of Respiratory and Critical Care Medicine, 2016. **193**(4): p. 427-437.
17. Shen, J., et al., *Analysis of MicroRNAs in Sputum to Improve Computed Tomography for Lung Cancer Diagnosis.* Journal of Thoracic Oncology, 2014. **9**(1): p. 33-40.
18. Doi, K., *Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential.* Computerized Medical Imaging and Graphics, 2007. **31**(4-5): p. 198-211.
19. Matsuki, Y., et al., *Usefulness of an Artificial Neural Network for Differentiating Benign from Malignant Pulmonary Nodules on High-Resolution CT: Evaluation with Receiver Operating Characteristic Analysis.* American Journal of Roentgenology, 2002. **178**: p. 657-663.

20. Gomez, M., et al., *Development and Testing of Multivariate Statistical Model To Predict Malignancy of Small (<1.5cm) Pulmonary Nodules.* American Journal of Respiratory and Critical Care Medicine, 2009. **179**: p. A1109.
21. Sun, T., et al., *Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set.* Computer Methods and Programs in Biomedicine, 2013. **111**(2): p. 519-24.
22. Zinovev, D., et al., *Predicting Radiological Panel Opinions Using a Panel of Machine Learning Classifiers.* Algorithms, 2009. **2**(4): p. 1473-1502.
23. Chen Hui, W.X.-h., M. Da-quing, and M. Bin-rong, *Neural network-based computer-aided diagnosis in distinguishing malignant from benign solitary pumonary nodules by computed tomography.* Chinese Medical Journal, 2007. **120**(12): p. 1211-1215.
24. Shah, S., et al., *Computer-aided lung nodule diagnosis using a simple classifier.* International Congress Series, 2004. **1268**: p. 952-955.
25. Lo, S.C.B., et al., *Classification of Lung Nodules in Diagnostic CT: An Approach Based on 3-D Vascular Features, Nodule Density Distributions, and Shape Features.* Proceedings of SPIE Medical Imaging Conference, 2003. **5032**: p. 183-189.
26. Li, Y. and J. Wang, *A mathematical model for predicting malignancy of solitary pulmonary nodules.* World Journal of Surgery, 2012. **36**(4): p. 830-5.
27. Swensen, S., et al., *The Probability of Malignancy in Solitary Pulmonary Nodules.* Archives of Internal Medicine, 1997. **157**(8): p. 849-855.
28. Takashima, S., et al., *Indeterminate Solitary Pulmonary Nodules Revealed at Population-Based CT Screening of the Lung: Using First Follow-up Diagnostic CT to Differentiate Benign and Malignant Lesions.* American Journal of Roentgenology, 2003. **180**: p. 1255-1263.
29. McNitt-Gray, M.F., et al., *A pattern classification approach to characterizing solitary pulmonary nodes inaged on high resolution CT: Preliminary results.* Medical Physics, 1999. **26**(6): p. 881-888.
30. McNitt-Gray, M.F., et al., *The effects of co-occurrence matrix based texture parameters on the classification of solitary pulmonary nodules imaged on computed tomography.* Computerized Medical Imaging and Graphics, 1999. **23**: p. 339-348.
31. Armato, S.G., et al., *Automated lung nodule classification following automated nodule detection on CT: A serial approach.* Medical Physics, 2003. **30**(6): p. 1188-1197.
32. Takashima, S., et al., *Small Solitary Pulmonary Nodules (>= 1 cm) Detected at Population-Based CT Screening for Lung Center: Reliable High-Resolution CT Features of Benign Lesions.* American Journal of Roentgenology, 2003. **180**: p. 955-964.
33. Gurney, J.W., *Determining the Likelihood of Malignancy in Solitary Pulmonary Nodules with Bayesian Analysis: Part I. Theory.* Radiology, 1993. **186**: p. 405-413.
34. Gurney, J.W., D.M. Lyddon, and J.A. McKay, *Determining the Likelihood of Malignancy in Solitary Pulmonary Nodules with Bayesian Analysis. Part II. Application.* Radiology, 1993. **186**: p. 415-422.
35. Aoyama, M., et al., *Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose CT images.* Medical Physics, 2003. **30**(3): p. 387-394.
36. Way, T.W., et al., *Computer-aided diagnosis of pulmonary nodules on CT scans: Improvement of classification performance with nodule surface features.* Medical Physics, 2009. **36**(7): p. 3086-3098.

37. Lin, P.-L., et al., *Automatic classification for solitary pulmonary nodule in CT image by fractal analysis based on fractional Brownian motion model.* Pattern Recognition, 2013. **46**(12): p. 3279-3287.
38. Hua, K.-L., et al., *Computer-aided classification of lung nodules on computed tomography images via deep learning technique.* OncoTargets and therapy, 2015. **8**.
39. COPDGene, *COPDGene COPD Genetic Epidemiology*.
40. Longanathan, R.S., et al., *Prevalence of COPD in Women Compared to Men Around the Time of Diagnosis of Primary Lung Cancer.* Chest, 2006. **129**(5): p. 1305-1312.
41. Eisenhauer, E.A., et al., *New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1).* European Journal of Cancer, 2009. **45**(2): p. 228-47.
42. Regan, E.A., et al., *Genetic Epidemiology of COPD (COPDGene) Study Design.* COPD, 2010. **7**(1): p. 32-43.
43. Laws, K.I., *Rapid texture identification,* in *Image Processing for Missile Guidance*, T.F. Wiener, Editor 1980, The Society of Photo-Optical Instrumentation Engineers: Bellingham, Washington. p. 376-280.
44. Sahiner, B., et al., *Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis.* Medical Physics, 1998. **25**(4): p. 516-526.
45. Jarvis, S., *Data mining with learner corpora*, in *A Taste for Corpora*, F. Meunier, et al., Editors. 2011, John Benjamins Publishing Company: Philadelphia, PA. p. 127-151.
46. DeLong, E.R., D.M. DeLong, and D.L. Clarke-Pearson, *Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.* Biometrics, 1988: p. 837-845.
47. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization paths for generalized linear models via coordinate descent.* Journal of statistical software, 2010. **33**(1): p. 1.
48. Freund, Y., R. Schapire, and N. Abe, *A short introduction to boosting.* Journal-Japanese Society For Artificial Intelligence, 1999. **14**(771-780): p. 1612.